F.YILMAZ

METU 2020

AN INVESTIGATION OF THE PSYCHOMETRIC PROPERTIES OF A LANGUAGE ASSESSMENT LITERACY MEASURE

FAHRİ YILMAZ

JULY 2020

AN INVESTIGATION OF THE PSYCHOMETRIC PROPERTIES OF A LANGUAGE ASSESSMENT LITERACY MEASURE

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF SOCIAL SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

FAHRİ YILMAZ

IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF ARTS
IN
THE DEPARTMENT OF ENGLISH LANGUAGE TEACHING

JULY 2020

Approval of the Graduate School of Social Sciences

____

Prof. Dr. Yaşar Kondakçı
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Arts.

____

Prof. Dr. Çiğdem Sağın Şimşek
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Arts.

____

Prof. Dr. Çiğdem Sağın Şimşek
Supervisor

Examining Committee Members

Prof. Dr. Kemal Sinan Özmen (Gazi Uni., ELT) __________________________

Prof. Dr. Çiğdem Sağın Şimşek (METU, FLE) __________________________

Prof. Dr. Bilal Kırkıcı (METU, FLE) __________________________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name, Last name	: Fahri Yılmaz
Signature	:

ABSTRACT

AN INVESTIGATION OF THE PSYCHOMETRIC PROPERTIES OF A LANGUAGE ASSESSMENT LITERACY MEASURE

Yılmaz, Fahri
Department of English Language Teaching
Supervisor : Prof. Dr. Çiğdem Sağın Şimşek

July 2020, 180 pages

This study investigates the psychometric properties of a modified measure designed to assess the knowledge base of EFL teachers' assessment literacy (AL). Using the data obtained from a sample of 4th grade pre-service EFL teachers from two state universities in Ankara, the psychometric properties of the measure were analysed by making use of several CTT-based and IRT-based analytical techniques. The findings indicate a good model fit, a presence of validity and high levels of reliability. Analyses of the sample's performance suggest that the measure was found to have a moderate difficulty level for the sample group, who exhibited a lower-than-expected level of achievement on the measure, and that CGPA was the only variable to statistically and positively correlate with the AL score. These findings point towards several important psychometric and pedagogical implications.

Keywords: Assessment Literacy, Language Assessment Literacy, Foreign Language Education, Assessing Assessment Literacy

ÖZ

DİLDE ÖLÇME-DEĞERLENDİRME OKURYAZARLIĞINA YÖNELİK BİR ÖLÇEĞİN PSİKOMETRİK ÖZELLİKLERİNE DAİR BİR İNCELEME

Yılmaz, Fahri
Yüksek Lisans, İngiliz Dili Eğitimi Bölümü
Tez Danışmanı : Prof. Dr. Çiğdem Sağın Şimşek

Temmuz 2020, 180 sayfa

Bu çalışmada İngiliz Dili Eğitimi öğretmenlerinin ölçme-değerlendirme okuryazarlığının bilgi temelini ölçmeyi hedefleyen uyarlanmış bir ölçeğin psikometrik özelliklerinin incelenmesi amaçlanmıştır. Ankara'da bulunan iki devlet üniversitesinde öğrenimlerini sürdürmekte olan 4. sınıf İngilizce öğretmeni adaylarından oluşan bir örneklem grubundan elde edilen verilerin incelendiği bu çalışmada, çeşitli Klasik Test Kuramı ve Madde Tepki Kuramı temelli analiz teknikleri kullanılmıştır. Araştırmanın sonuçları iyi bir model uyumluluğuna, geçerliğe ve yüksek düzeyde güvenirliğe işaret etmektedir. Örneklem grubunun performans analizi, ölçeğin kitle tarafından orta güçlük düzeyinde bulunduğunu ve kitlenin kendilerinden beklenenin altında bir başarı gösterdiğini ve bu başarıyla istatistiki ve pozitif ilişkisi olan tek değişkenin ağırlıklı genel not ortalaması olduğunu ortaya koymuştur. Bu bulgulara dayanarak birtakım önemli psikometrik ve pedagojik çıkarımlara varılmıştır.

Anahtar Kelimeler: (Dilde) Ölçme-değerlendirme Okuryazarlığı, Yabancı Dil Eğitimi, Ölçme-değerlendirme Okuryazarlığının Ölçülmesi

To my wife, and best friend, Dilek Yılmaz

ACKNOWLEDGMENTS

A number of great professors in the English Language Teaching program at METU have contributed to my professional and academic development during my MA study. I am indebted to them for their guidance, encouragement and support.

I am especially grateful to my supervisor Prof. Dr. Çiğdem Sağın Şimşek, who has been a great source of knowledge and help, for her complete and accessible support whenever I needed it. I have benefited enormously from her knowledge throughout my study.

I owe very special thanks to Assist. Prof. Dr. Semirhan Gökçe, who patiently read the methodology and findings chapters of this study, and provided invaluable feedback.

I thank Prof. Dr. Kemal Sinan Özmen for not only providing me with constructive feedback and great ideas, but also for being a great source of inspiration for me throughout all these years we have known each other.

I wish to extend my sincere thanks to all participating pre-service teachers at Gazi University and METU. This study would not have been possible without them.

I am more than grateful to my beloved wife, parents and friends for their support and patience, and I would like to convey my apologies to them for stealing from their time. Also, I must thank our cat Turşu (Pickle) for revealing his constant presence by my side (on my desk) and helping alleviate my stress levels during long study nights by allowing me to pat him.

Last but not the least, I wish to extend my most sincere thanks to health professionals in Turkey and around the world for all their incredible efforts to keep all of us safe and sound, as the world is going through one of the hardest pandemics in our history due to Covid-19.

TABLE OF CONTENTS

PLAGIARISM. iii

ABSTRACT iv

ÖZ v

DEDICATION vi

ACKNOWLEDGMENTS vii

LIST OF TABLES xii

LIST OF FIGURES xiii

LIST OF ABBREVIATONS xiv

CHAPTER

1. INTRODUCTION 1

  1.1 Assessment Literacy 1

  1.2 Language Assessment Literacy 4

  1.3 Rationale and Research Questions 6

  1.4 Significance of the Study 8

2. LITERATURE REVIEW 9

  2.1 Second Language Teacher Education 9

  2.2 Fundamental Considerations in Assessment 10

    2.2.1 Basic Concepts in Assessment 10

      2.2.1.1 Assessment 10

      2.2.1.2 Measurement 12

      2.2.1.3 Test 12

      2.2.1.4 Evaluation 13

    2.2.2 Types of Assessment 13

      2.2.2.1 Purpose 13

      2.2.2.2 Other Types of Assessment 15

        2.2.2.2.1 Norm-referenced and criterion-referenced tests 15

2.2.2.2.2 Summative vs. formative assessment 16
2.2.2.2.3 Objective vs. subjective testing 17
2.2.2.2.4 Direct vs. indirect testing 17
2.2.2.2.5 Discrete-point vs. integrated tests...18
2.2.3 Qualities of a Good Test 19
2.2.3.1 Practicality 19
2.2.3.2 Reliability 19
2.2.3.3 Validity...21
2.2.3.4 Scoring...25
2.3 Assessment and Teaching 27
2.4 Assessment Literacy 28
2.4.1 Definition of AL...28
2.4.1.1 Frameworks for AL..34
2.4.1.2 Measurement of Assessment Literacy...38
2.4.2 Significance of AL 39
2.5 Language Assessment Literacy 41
2.5.1 Definition of LAL...41
2.5.2 Findings on EFL Teachers' LAL 42
3. METHODOLOGY 44
3.1 Purpose..44
3.2 Research Questions...45
3.3 Context and Participants 45
3.4 Data Collection Procedures...49
3.4.1 Data Collection Instrument 49
3.4.2 Administration of the Measure...51
3.4.3 Development and Adaptation of the Measure..51
3.5 Data Analysis...55
3.5.1 Rasch Analysis 56
3.5.2 CTT Methodology 57
3.5.3 IRT...59
3.5.4 Rasch PCA 60
4. FINDINGS AND DISCUSSION 62

4.1 Findings 62
4.1.1 1PL Model 62
4.1.2 CTT Methodology 68
4.1.3 Rasch Analysis 73
4.1.4 2PL Model 77
4.1.5 Rasch PCA 82
4.1.6 Item Difficulties/Person Performance 86
4.1.7 Correlations 89
4.2 Discussion 91
4.2.1 Research Question 1 91
4.2.2 Research Question 2 102
4.2.3 Research Question 3 106
5. CONCLUSION 109
5.1 Overview of the Study 109
5.1.1 Overview of Rationale of the Study 109
5.1.2 Overview of Methodology 110
5.1.2.1 Instrument (CAK) 110
5.1.2.2 Data Collection 111
5.1.2.3 Data Analysis 111
5.1.3 Overview of the Results 112
5.2 Implications 113
5.2.1 Psychometric Implications 113
5.2.2 Pedagogical Implications 114
5.3 Limitations and Future Directions 115
5.4 Conclusion 116
REFERENCES 118

APPENDICES

A. HUMAN SUBJECTS ETHICS COMMITTEE APPROVAL 136
B. DEBRIEFING FORM 137
C. INFORMED CONSENT FORM 138
D. ORIGINAL CAK 139
E. EVALUATION FORM FOR MODIFIED CAK 146

F. MODIFIED CAK 150
G. TURKISH SUMMARY/TÜRKÇE ÖZET..161
H. TEZ İZİN FORMU / THESIS PERMISSION FORM 180

LIST OF TABLES

Table 1: Popham’s (2009) Suggested Content Points for Teacher AL 29

Table 2: Standards for Teacher Competence in Educational Assessment.............. 31

Table 3: Summary of Studies Using Some of the Objective Measures to Directly Test Teacher Assessment Knowledge....................................... 40

Table 4: Participant Descriptive Statistics (1).. 48

Table 5: Participant Descriptive Statistics (2).. 49

Table 6: Reordering of the Items on the Measure... 54

Table 7: Items on the Modified Measure and Corresponding Standards 55

Table 8: Analytical Techniques Used in Data Analysis.. 56

Table 9: An Overview of the Item difficulty Values from 1PL Analysis 63

Table 10: Descriptive Statistics from the Traditional Item Analysis (N=74)......... 69

Table 11: Results from the Traditional Item Analyses (N=74) 69

Table 12: Options Analysis for the Eliminated Items 71

Table 13: Contents of the Eliminated Items 72

Table 14: Summary of Rasch Person and Item Statistics...................................... 74

Table 15: Rasch Analysis Item Statistics.. 75

Table 16: 2PL Goodness of Fit Analysis.. 78

Table 17: Results from the 2PL Analysis... 80

Table 18: Modified CAK Item Loadings (N=74) 84

Table 19: Items Ranked by Difficulty.. 87

Table 20: Comparison of Participant Achievement by Standards 88

Table 21: Descriptive Statistics for CGPA and Total Score.................................. 91

Table 22: Clustering of Items and Standards into Two Components................... 100

Table 23: Ordering of Standards by Difficulty 104

LIST OF FIGURES

Figure 1: The Standards and Corresponding Items on the Original Measure 53

Figure 2: An Example ICC (Szabó, 2012) 59

Figure 3: ICCs for All Items...66

Figure 4: Item Information Curves for All Items..67

Figure 5: Test Information Function (1PL) 67

Figure 6: Item-person Map (N=74)...77

Figure 7: Scatter Plot for Item Difficulty and Discrimination.................................81

Figure 8: Item Characteristic Curves (ICCs)..82

Figure 9: Test Information Function (2PL) 82

Figure 10: Rasch PCA Variance and Components.. 83

Figure 11: Scatter Plot Matrix 90

Figure 12: ICCs Belonging to Q18, Q19, and Q21...99

LIST OF ABBREVIATONS

1PL	One-parameter Logistic Model
2PL	Two-parameter Logistic Model
3PL	Three-parameter Logistic Model
AL	Assessment Literacy
ALI	Assessment Literacy Inventory
AFT	American Federation of Teachers
CAK	Classroom Assessment Knowledge Instrument
CTT	Classical Test Theory
EFL	English as a Foreign Language
ICC	Item Characteristic Curve
IIC	Item Information Curve
IRT	Item Response Theory
LAL	Language Assessment Literacy
NCME	National Council on Measurement in Education
NEA	National Education Association

CHAPTER 1

INTRODUCTION

1.1 Assessment Literacy

Stiggins (1991) defines assessment literacy (AL) as a basic understanding of educational assessment and the skills related to it. It has been increasingly recognised that AL is an essential skill teachers need to possess (Popham, 2009; Xu & Brown, 2016). There is a wide consensus that teachers with a thorough understanding of assessment can make sophisticated and informed decisions and judgments about the validity and reliability of practices and policies related to assessment in a variety of contexts. On the other hand, a teacher whose AL level is insufficient may end up carrying out assessment practices that are not valid and reliable, hence misinforming not only the students but also other stakeholders including parents, other teachers, and school administration. Therefore it would be safe to state that teacher AL is closely related to the success of both educational assessment and quality of education in general.

Teachers can be empowered with AL as it can help them reach informed decisions when developing, administering and using assessments (Harding & Kremmel, 2016), whereas teachers without sufficient AL may end up leading students to suffer adverse consequences including failing to be advanced to the next level despite deserving it, and failing to receive additional support despite needing it (Purpura, 2016; Purpura, Brown & Schoonen, 2015). Cheng (2001) informs that up to a third of teachers’ time is allocated to activities related to testing; however, most teachers have little or no training to carry out the assessment-related activities (Bachman, 2000). A similar view is held by Coombe, Troudi and Al-Hamly (2012), who consider that teachers cannot provide students with the necessary support in terms of obtaining higher levels of academic achievement if they do not possess a high level of AL.

The role AL plays in teaching and learning processes is quite important. According to White (2009), learning can be initiated by assessment, which can be considered like a locomotive. Assessment cannot be considered separate from learning and teaching processes because teachers are constantly involved in assessment-related activities whether they are formal or informal, or traditional or alternative assessment activities, which makes AL or good assessment skills significant for teachers in order to ensure the quality of teaching and learning (Stiggins, 1991). Teachers are expected to be equipped with the insight provided by assessment-related activities for a large number of educational purposes including identifying whether course content is relevant or not, enhancing the learning and teaching processes, the effectiveness and efficacy of the instruction, and informing learners on their current ability or achievement levels as well as their strengths and weaknesses in relation to the expected learning outcomes of a course (Mertler, 2003). According to the author, the profession of a teacher requires him or her to take the assessment responsibility. Moreover, because teaching and assessment constantly provide each other with information that can be used to improve both (Malone, 2013), teachers are expected to bridge the two educational concepts. As highlighted by several researchers (Stiggins, 1999; Popham, 2009), teachers who are equipped with sound knowledge and mastery of the concept of assessment can make more informed decisions in their profession, which can have a big effect on the quality of education (Malone, 2013).

Teachers who are literate in assessment are teachers who know "what they are assessing, why they are doing it, how best to assess the skills, 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" (Stiggins, 1995, p. 240). This implies that AL is related not only to assessment knowledge but also to the application of this knowledge to assessment practices.

It is agreed by assessment researchers that a good understanding of both classroom assessment and large-scale assessment requires the use of cognition, observation and interpretation (National Research Council, 2001). In other words, these three concepts constitute the backbone of any assessment system as they afford evidence for sound validation efforts in order to ensure fair and appropriate uses of assessment data. The model provided here by the National Research Council points to these three components in defining AL, where cognition is related to a teacher's understanding of student cognition, observation is related to a teacher's understanding of assessment tools, and interpretation refers to a teacher's understanding of data interpretation. A competent teacher, therefore, is expected to be able to carry out assessment-related practices in a systematic and evidence-based way, and make use of the insight provided by the increasing research area.

As the interest in AL has intensified for the past several years with the recognition of AL as an important component of teacher professional development programs (Beziat & Coleman, 2015), an increased presence of concepts is observed related to educational assessment in pre-service and in-service programs (Mertler, 2003; Alkharusi, Kazem, & Al-Musawai, 2011; Xu & Brown, 2016). According to Stiggins (2006), teachers and instructors in US schools and universities have unacceptably low levels of AL, which leads to inaccuracy in assessing learners' abilities and learners' failure to achieve their full potential. The fact that many teachers graduate from their undergraduate programs inadequately equipped with AL forces them to obtain AL skills on the job (Mertler, 2003). Also, many teachers who have acceptable levels of knowledge in classroom assessment lack the knowledge or skills needed to interpret data provided by large-scale or high-stakes exams (Conor & Mbaye, 2002). Stiggins (2006) informs that such exams are provided by authorities in the educational systems, and teachers with no control over the content of these tests, are compelled to teach for these tests (Xu & Brown, 2016). This lack of knowledge and interpretation skills leaves teachers unprepared to use valid procedures of evaluation (Yan & Cheng, 2015).

Even though there has been an ever more significant emphasis on AL in pre-service and in-service programs, research finds insufficiencies in both classroom assessment literacy and large-scale assessment literacy among teachers (Mellati, Khademi, & Shirzad, 2015). The evidence from many countries suggests that there are a large number of teachers who lack adequate training and knowledge in the development, administration, and interpretation of different assessment tools. Teachers demonstrate this lack of knowledge not only in common assessment responsibilities but also in the understanding of the basic concepts of assessment such as validity and reliability (Gotch, 2012). For instance, research has shown that rubrics created by many teachers are of average quality, far from reflecting the best and up-to-date practices or making clear links between instruction and assessment (Maclellan, 2004). Several teachers have been found to self-assess their AL to be high; however, studies have found that even teachers with essential assessment skills may have difficulty with such assessment activities as test construction, which they consider to be complex (Al-Maliki & Weir, 2014; Scott, Webber, Aitken & Lupart, 2011).

Research suggests that teachers without adequate skills in developing strong assessment tools find it difficult to engage with new types of assessment tools as opposed to more conventional pen-and-paper exams (Wiliam & Thompson, 2008). There is also a constant gap between assessment practices and instructional goals. Similarly, research carried out by Susuwele-Banda (2005) found that teachers paid more attention to measuring the learners' mastery, and outcomes, and that they frequently used performance-based evaluation. The teachers in the study, who were interested in measuring learner achievement, also considered classroom assessment as an essential practice for their teaching, but not for improving their teaching. However, they were found to be lacking the skills and insight needed to understand and analyse the reasoning behind the responses provided by their students. On the other hand, there is also substantial research with opposing results. Several researchers (Dayal & Lingam, 2015; Gotch, 2012; Dinther, Dochy, & Segers, 2015) found that teachers do not like tests as they believed that tests result in unnecessary stress and exhaustion for learners, which could explain why teachers are generally found by research not to be good at judging the quality of their own assessment practices as well as evaluating their students' ability (Bastian, Henry, Pand, & Lys, 2016; Clark-Gareca, 2016).

1.2 Language Assessment Literacy

Recently, the concept of language assessment literacy (LAL) has emerged, which originated from the literature in AL, but it can be considered to be distinct from AL in general for a number of reasons. There are various definitions of LAL. According to Malone (2013, p. 329), LAL relates to "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". Inbar-Lourie (2008, pp. 389-390) defines language assessment knowledge as a base comprising "layers of assessment literacy skills combined with language-specific competencies, forming a distinct entity that can be referred to as language assessment literacy”. She also adds in another work (2017) that the term LAL stemmed from AL, yet it is distinct from AL because it endeavours to “incorporate unique aspects inherent in theorising and assessing language-related performance” (p. 259). These definitions highlight the ‘language-specific’ aspect of LAL, which sets it apart from AL, while it draws on the literature and principles of AL. In other words, LAL addresses additional skills related to the nature of language as compared to AL. Nevertheless, according to Fulcher, LAL “is still in its infancy” (2012, p. 117).

Price, Rust, O’Donovan, Handley, and Bryant (2012) underlined the need for language educators to be adequately knowledgeable in assessment-related procedures. Yet, research suggests that many teachers lack the assessment knowledge needed (Plake, 1993). Stiggins (2010) referred to this problem as ‘language assessment illiteracy’ that thrives among teachers (p. 233). It demonstrates that although teachers are expected to have LAL skills, how assessment-literate they are is controversial. According to Xu and Brown (2017), AL begins with the knowledge base, and thus, knowledge of assessment is central to AL.

However, several research studies investigating EFL teachers’ LAL have shown that teachers have problems with understanding even the basic principles of LAL, or with applying them in their practices. For instance, Lam (2015), who aimed to find out about whether two language assessment courses contributed to LAL of pre-service teachers in five institutions in Hong Kong, found that there was not sufficient support to enhance LAL, and the training was inadequate. Tsagari and Vogt (2017) also wanted to explore in-service teachers’ perceptions of LAL, and they found that the perceived LAL of participants from institutions in Cyprus, Greece, and Germany was inadequate.

A review of studies of LAL demonstrates a number of problems experienced by EFL teachers in terms of language assessment knowledge. López and Bernal (2009) conducted a research study that indicates a presence of different assessment practices among EFL teachers. For instance, teachers who have training in LAL often use assessment with the purpose of enhancing teaching and learning (for formative purposes), whereas teachers who have little or no training in language assessment use assessment only to obtain grades from learners (for summative purposes), which indicates, according to the authors, that the teachers who lack LAL make no distinction between types of assessment and grades.

López and Bernal (2009) carried out their research in Colombia. However, research findings coming from other parts of the world including Chile (Díaz, Alarcon, & Ortiz, 2012), China (Cheng, Rogers, & Hu, 2004), and Canada (Volante & Fazio, 2007) resonates with the findings of López and Bernal, which suggests that there is a need for EFL teachers to improve their assessment practices to enhance both teaching and learning. Even though there are not many research studies exploring the LAL levels of EFL teachers in the Turkish context to the knowledge of the researcher, there are several studies so far in Turkey including Hatipoğlu (2015, 2017), Mede and Atay (2017) Ölmezer-Öztürk and Aydın (2018), Öz and Atay (2017), and Şahin (2019), which have found that EFL teachers in Turkey exhibit low levels of LAL.

1.3 Rationale and Research Questions

Teachers, whether at primary, secondary or tertiary level, are often tasked with designing, developing, and/or choosing assessment methods, administering assessment tools, using assessment results to provide feedback, scoring and grading, recording information obtained from assessment, and reporting assessment results to key stakeholders, including but not limited to students, school and ministry administrators, parents, potential employers and other teachers (Lamprianou & Athanasou, 2009; McMillan, 2014; Popham, 2014; Russell & Airasian, 2012; Taylor & Nolen, 2008). These assessment-related activities take up one-third to half of teachers' instructional time (Bachman, 2014; Mertler, 2003; Stiggins, 1991, 1995), which emphasises the idea that the quality of teaching and student learning could be directly related to the quality of assessment practices undertaken by teachers in the classroom (Earl, 2013; Green, 2013). For this reason, teachers are expected to establish a congruent mediation procedure between their assessment and instruction practices in a way that would enhance student learning (Earl, 2013; Griffin, Care, & McGaw, 2012; Popham, 2014; Shepard, 2008). Such an approach might offer the opportunity to equip learners with twenty-first century skills including lifelong learning, which involves subskills like critical-thinking, problem-solving, creativity, flexibility and cultural appreciation. Accordingly, teachers are expected to possess the knowledge of and skills related with assessment in order to be able to design, choose and administer assessment tasks tailored for learner needs through a shift from a testing culture to an assessment culture (Masters, 2013).

Many assessment researchers have stated that meeting the goal of equipping learners with the twenty-first century skills requires teachers to be able to make use of a wide range of assessment methods in assessing student learning for both formative and summative purposes (Black & William, 1998a, 1998b; Griffin et al., 2012, Heritage, 2013; Masters, 2013; Shute, 2008). Among such methods are portfolios, performance-based tasks, and peer and self-assessment in addition to the use of more traditional assessment tools. Proper use of assessment instruments and assessment results to enhance and improve instruction and learning as well as supporting lifelong learning come with numerous benefits including improvement of higher-order thinking skills (Darling-Hammond & Adamson, 2013; Leighton, 2011; Moss & Brookhart, 2012), enhancing student motivation for learning, helping students become autonomous learners and become owners of their own learning (Falchikov & Boud, 2008; Heritage, 2013; Lamprianou & Athanasou, 2009; Molloy & Boud, 2014; Nicol, 2013).

Even though possessing the knowledge of and skills associated with high quality educational assessment generates a number of benefits, researchers have continually reported findings indicating poor AL and poor assessment practices among teachers both in the wider field of education (Plake, 1993; Stiggins, 2010) and in language education (Tsagari & Vogt, 2017; Volante & Fazio, 2007; Xu & Brown, 2017). Research has shown that lack of understanding of assessment and presence of poor assessment practices may lead to a mismatch between assessment and instruction/learning goals (Binkley et al., 2012; Griffin et al., 2012; Heritage, 2013; Rea-Dickins, 2007). The mismatch between the importance of high quality assessment and teachers' poor assessment knowledge and skills in addition to inadequate emphasis by pre- and in-service education on assessment leads to the problematizing of LAL (Stabler-Havener, 2018), which creates the need to discuss the question of whether it is possible to measure EFL teachers' LAL. Closely related to this question are follow-up questions of how pre-service EFL teachers at two higher education settings in Turkey perform on an assessment instrument that tests their LAL at the knowledge base, and what factors affect their LAL. Therefore, the current study endeavours to gain insights into LAL of pre-service EFL teachers specifically focussing on these issues. The following paragraphs present a formulation of the current study's research questions based on the research objectives as well as providing an overview of the significance of the study.

Research Question 1: What are the psychometric properties of the adapted Classroom Assessment Knowledge instrument, devised to assess EFL teachers' language assessment literacy knowledge base?

Research Question 2: What is the language assessment literacy knowledge base level of pre-service EFL teachers in the higher education context in Turkey?

Research Question 3: What factors, if any, affect language assessment literacy of pre-service EFL teachers in the higher education context in Turkey?

1.4 Significance of the Study

To the knowledge of the researcher, this is the first empirical research study in Turkey into the psychometric properties of a LAL measure based on a widely recognised AL framework, modified and contextualised to the Turkish context to assess EFL teachers' assessment knowledge. With AL gaining attention and importance both in language education and the broader field of education, the need to accurately assess the assessment knowledge of teachers and pre-service teachers is becoming an important issue. Although a complete assessment and understanding of teachers' AL requires a complete evaluation of both their knowledge and practice, the knowledge base of AL is an important indicator of the wider AL of teachers including their ability to put the knowledge into practice. Therefore, this study aims to contribute to the literature by examining the psychometric properties (i.e., validity and reliability) of a potential LAL measure. The results of the current study could also be used to obtain some insight into the current LAL levels of pre-service EFL teachers in Turkey to inform policies and decisions regarding their needs, strengths and weaknesses.

CHAPTER 2

LITERATURE REVIEW

This chapter presents an overview of the existing body of literature relevant to the research objectives of the current study. The following paragraphs discuss (a) an overview of second language teacher education (SLTE) and SLTE in Turkey, (b) some of the most important topics in educational assessment, (c) the relationship between assessment and teaching as well as the importance of assessment in teaching, (d) definition and significance of AL, and (e) definition of LAL and important findings of several research studies on EFL teachers' AL.

2.1 Second Language Teacher Education

English is accepted as the "global language" (Crystal, 2003, p. 1), and it gained this status thanks to being an official language of many countries and by being the language primarily taught as a second language around the world. English is becoming a compulsory school subject in many countries at younger and younger ages (Nunan, 2001). As the number of English language learners increase around the world rapidly, the demand for English language teachers is becoming unavoidable, which brings SLTE to the forefront (Bailey, 2001; Wright, 2010). The terms "teacher education" and teacher training" are used in the literature to address this demand (Freeman, 2001). Even though Widdowson (1997) makes a distinction between the two concepts, according to which teacher education focuses more on practical terms (solution-oriented) while teacher training is more problem-oriented and focusses on theoretical considerations, the two concepts are often used interchangeably. Embracing both concepts, Richards and Nunan (1990) describes the aim of SLTE as "to provide opportunities for the novice to acquire the skills and competencies of effective teachers and to discover the working rules that effective teachers use" (p. 15).

As the field of SLTE has gone through a number of theoretical and practical developments in the effort to train EFL teachers, the concept of professionalism has gained substantial importance, and as Richards (2008) informs, “becoming an English teacher means becoming part of a worldwide community of professionals with shared goals” (p. 161). Similar to the developments in the broader field of education, one important consequence of the increasing professionalism in SLTE has been the creation of standards that have become popular in the field (Richards, 2008).

Parallel to the global policies, substantial significance has been placed on English in Turkey as well, and English has become the only foreign language as a compulsory subject at all educational levels (Kırkgöz, 2009). Öztürk and Atay (2010) describe the role of English in the Turkish educational system as follows:

> Today English education is offered from kindergarten level until university, either as a compulsory foreign language or as the means of instruction, e.g., there are many secondary schools and universities with a one-year preparatory class followed by English-medium instruction. In addition to the private English courses, the government encourages citizens of all ages to become proficient in English by expanding educational opportunities (p. 137)

2.2 Fundamental Considerations in Assessment

Fundamental considerations in educational assessment will be reviewed in this section in three categories: (a) basic concepts in assessment, (b) types of assessment, and (c) qualities of a good test.

2.2.1 Basic Concepts in Assessment

Among the most important and basic concepts in the field of assessment are assessment, measurement, testing and evaluation. Even though these terms are frequently used with different meanings and often interchangeably, there is a need to understand the nuances of these terms.

2.2.1.1 Assessment

A range of different meanings have been attributed to the term ‘assessment’ in educational sciences, and different researchers in the field of educational measurement and language assessment have used the term in various ways, which suggests that no consensus exists over what exactly it means (Bachman, 2014, p. 7). In addition, several terms including “test(ing)”, “measurement”, and ‘evaluation” are often used interchangeably to refer to assessment. However, despite the wide variety of meanings assigned, it is generally agreed that assessment refers to the process of gathering information regarding an object of interest using “systematic’ and “substantively grounded” procedures, and except for cases where the object of interest is student information such as attitudes or demographic characteristics, the object of interest of a language assessment activity is one aspect or a combination of aspects of language ability. The term ‘assessment’ is also frequently used to refer to the product of this information-gathering process.

Bachman (2014) also provides a clear summary of the two properties of assessment: being systematic and being substantively grounded. These two properties distinguish language assessment from informal observation. Being systematic means that the design and implementation of the assessment are described clearly, allowing other individuals to reproduce it if they wish to do so. Systematicity is closely related to the principle of reliability, which will be discussed later in this section.

The other property is being substantively grounded. It is related to forming a basis for the interpretation of both quantitative and qualitative results of an assessment. It must be a widely-accepted theory about the nature of language, language ability, language use, language learning, or previous research as well as acknowledged practice that forms the basis of language assessment. This property is closely related to the principle of validity, which will also be discussed later in this section.

According to Chan (2008, p.7) “assessment refers to any method, strategy or tool a teacher may use to collect evidence about students’ progress toward the achievement of established goals”. In assessment, the information collected and the evidence gathered help to understand what students have learned. Heaton (1990) summarizes that assessment aims to (a) have an understanding of the students’ strengths and weaknesses in learning, (b) helps teachers better understand and monitor the process of learning experienced by learners, (c) make evaluation about their learning, and (d) use the assessment and evaluation information in order to place learners in appropriate groups based on institutional standards. Teachers are expected to use assessment in several ways including making interpretations and decisions about their students’ learning, and enhancing their teaching by reflecting on the assessment practices and activities. It is worth noting that teachers can get useful and immediate feedback from assessment on what, how much, as well as how well learners are learning.

2.2.1.2 Measurement

Another fundamental concept in assessment is measurement. According to Bachman (1990), it refers to the "process of quantifying the characteristics of an object of interest according to explicit rules and procedures" (p. 18). And similar to assessment, measurement is also used to refer to a product or outcome of the process of measurement.

Measurement is a type of assessment that involves quantification, i.e.i assigning of numbers. This characteristic makes the distinction between measurement or measures and non-quantitative assessments like verbal descriptions or visual images. It is worth noting that numbers are assigned not directly to people, but to the attributes associated with individuals or groups of individuals. In language assessment, the attributes to be measured are usually not physical ones such as height or weight, but attributes or abilities that cannot be observed directly, including grammatical knowledge, communicative competence or language aptitude. Like in other assessment types, measurements are also administered and implemented based on explicit rules and procedures in a systematic way. This is usually achieved through test specifications, criteria, valid and reliable scoring procedures and explicit test administration procedures. Through the use of these explicitly defined processes and procedures, a link between the unobservable trait to be measured and the observable performance to be quantified is established.

2.2.1.3 Test

Coombe (2018) defines a test as "a set of tasks or activities intended to elicit samples of performance which can be marked or evaluated to provide feedback on a test taker's ability or knowledge" (p. 41). It can be stated that a test is a specific type of measurement used to elicit a specific performance sample which we associate with a specific unobservable trait. One important implication of this definition is that during test development, particular tasks and sets of tasks are designed to elicit certain samples of performance linked with certain traits or unobservable abilities. Coombe also mentions another meaning frequently associated with the term 'test'. It is often used to refer "to the activity of measuring samples of performance elicited by a test from a test taker" (p. 40). This process can provide information regarding the test taker's level of content and skill acquisition.

2.2.1.4 Evaluation

Evaluation is a term frequently associated with assessment. Evaluation, which can be considered to be one possible use of assessment, is related to arriving at value judgments and decisions. Educational programs usually attach considerable importance to evaluation in making such decisions as selection, placement, collecting information about the worth of a program, and grading or marking.

Coombe (2018) mentions four levels of evaluation, especially when the term is used to refer to the process of using the results of an assessment to judge and support learning and instruction. These four levels are learner feedback, learner learning, learner behaviour, and learning results.

2.2.2 Types of Assessment

2.2.2.1 Purpose

Numerous specific types of assessment purposes can be divided into two general categories (Green, 2013). The first category relates to language learning, and it involves assessing to what extent a learning goal has been achieved. This type of purpose is often used in schools and other educational settings. The main focus is usually on what has been taught or will be taught, and these kinds of tests are usually designed and implemented by teachers. They are often flexible enough to allow teachers to use observational techniques such as watching and recording, portfolios (long-term collections of the work of the learners), self-assessment, and both informal tests and quizzes and formal tests carried out with more strictly-controlled conditions in place.

The second category of purposes relates to gathering information about an individual's language ability in general in order to understand whether their language ability satisfies a set of predetermined criteria or standards, which is referred to as proficiency assessment. It is usually linked to carrying out an assessment of language and related skills needed to perform a certain task such as carrying out a job, and studying an academic subject. As opposed to the aforementioned first purpose type, where the focus is on what content has been taught or will be taught, the focus of this type is not on what content a course or program has taught. This type of assessment is not likely to be developed or delivered by teachers but assessment professionals administering formal tests with controlled, standardised and uniform conditions and procedures in place. It is likely to be administered by professional national or private organisations.

The main distinction between proficiency assessment and educational assessment is that the former does not focus on specific learning processes or instruction outcomes (Carr, 2011). It is interested in finding out about the current functionality of an individual, not their learning process. Proficiency assessment seeks to understand what test-takers can do with their current language ability rather than how they have arrived at their current level. The key word in proficiency assessment is whether a test-taker can perform certain tasks or meet certain needs with his or her current ability.

Dividing assessments into certain categories or types is an arduous task. According to Brown (2004), language assessment is generally used to contribute to making certain decisions, and these decisions are needed based on various purposes. A broad categorisation can be made according to the purpose they are used for. In this context, Carr (2011) groups language tests into two main categorises based on the purposes they are used for: curriculum related decisions (admission, placement, diagnostic, progress, and achievement), and other decisions (proficiency and screening)

Admission test is the first type of curriculum-related test a new student may experience. It is used to determine if a student is eligible for being accepted into the program in the first place. Placement test, which is a related test to admission test, often goes hand in hand with admission test. It is used to determine a student level of study. It is often the case that one single test is used to serve both of the purposes, that is, not just to determine if a learner's language ability is sufficient for the program and to estimate the right level for him or her.

Learners' strengths and weakness areas are usually identified using diagnostic tests. Despite the fact that sometimes placement tests or admission tests in a language program may be used to identify learner needs, they are often designed and administered separately following the placement of students in the program. Diagnostic tests may also afford information regarding whether the placement has been carried out accurately, which is often preferred as a method by those programs that are not very confident about the quality of their placement test.

Teachers are expected to use the information obtained from diagnostic tests in order to design or refine their instruction based on the needs and strengths of learners.

After proper placement of learners, teachers may want to know if their students are learning what is being taught to them, or whether any learning takes place at all. It is through the use of progress tests that teachers assess the students' performance in terms of learning with respect to the learning outcomes of a course. As opposed to achievement tests, which are carried out to find out about to what extent students have satisfied or acquired the learning outcomes or objectives of a course, progress tests provide information about how well they are learning as they are delivered while the instruction or learning still takes place. Therefore, the distinction between a progress test and an achievement test, or the decision whether a test, or a quiz, is a progress test or an achievement test, is made in terms of how the results of the test or the quiz are being used.

2.2.2.2 Other Types of Assessment

Apart from the broader grouping explained in the previous paragraphs based on test purpose, Carr (2011, p.9) also proposes a categorisation of tests "in terms of frameworks for interpreting results, the things that examinees have to do during the test, and the ways that the tests are scored", several of which denote various dichotomies.

2.2.2.2.1 Norm-referenced and criterion-referenced tests

Norm-referenced and criterion-referenced testing represent two distinct frames of reference that help interpret the results or scores of a test. According to Thorndike and Hagen (1969), test score is only meaningful as long as it is compared to some reference. Whether the comparison is performed against other test-takers or against some predetermined standards or criteria defines the nature of this reference, which is what distinguishes norm-referenced testing from criterion-referenced testing.

In norm-referenced testing, the score of a test-taker is compared against the score, or the performance, of other test-takers who took the same test. The scores are often reported in terms of percentile scores, in other words, the percentage of other test-takers whose scores were lower than theirs. Due to the large numbers required to divide test-takers into groups of 100, it is natural that norm-referenced testing is usually used in large-scale testing, where testing professionals deal with large numbers of test-takers.

Because norm-referenced tests deal with the success of test-takers compared to that of other test-takers, they are far from informing the users of the test on how successful a test-taker is in absolute terms. Test administrators and other stakeholders of a test including test-takers, parents, and educational decision and policy makers usually demand to know more than provided by the norm-referenced tests, which makes criterion-referenced testing highly important. Such tests measure test-taker performance not in terms of a comparison of their performance against that of other test-takers, but in terms of a set of predetermined criteria and standards by looking at whether a test-taker successfully satisfies them. In criterion-referenced testing, test-taker scores are frequently reported in percentages rather than percentiles, that is, the percentage of the criteria satisfied by the tests-takers.

2.2.2.2.2 Summative vs. formative assessment

Summative vs. formative assessment is a way of looking at assessments in terms of an interpretation of assessment results based on when they are administered and for what purposes the results are used (Carr, 2011). If a test is administered at the end of a unit, program, course, etc., in order to collect information about to what extent students have learned the content, it is called a summative test, and it is often used for grading purposes.

Formative assessment, on the other hand, is the type of assessment given to learners while they are still in the process of learning in order to provide information about the quality of learning that is taking place (Bachman, 1990). By its nature, it is closely linked to progress assessment. The information obtained from formative assessment is usually used to help make decisions about whether there is a need for change in the course syllabus, instruction techniques, program, etc. However, although summative vs. formative assessment is usually perceived as some kind of a dichotomy, the distinction between the two types may not always be clear-cut, for the results obtained from a language quiz, for instance, may be used by a teacher both to provide revision information about the instruction and to assign grades.

2.2.2.2.3 Objective vs. subjective testing

Another dichotomy listed by Carr (2011) in the identification of test types is the “false distinction between objective and subjective testing” (p. 12). The term ‘objective test’ is usually used to refer to a test considered to be open to objective scoring that uses selected-response items such as multiple-choice questions, matching questions or true-false questions. However, such an approach is open to criticism and controversy because even the so-called objective tests involve subjectivity by their nature because those who decide the content, topics, and test specifications in general (such as the number of questions, length of passages, item types, etc.) of the test make subjective decisions. Subjective tests, however, are called subjective because they contain tasks that require human judgment for scoring. Yet, through the use of several well-established mechanism and practices such as introducing a valid and reliable scoring rubric, rater training, and robust statistical methods to increase interrater and intrarater reliability, such tests can be prevented from being as subjective (Carr, 2011).

2.2.2.2.4 Direct vs. indirect testing

According to Carr (2011), as in the case of objective vs. subjective testing, another problematic dichotomy lies in direct vs. indirect testing. What is often meant by direct tests are tests with items requiring test-takers to use the trait or ability that is intended to be assessed. For instance, it is called direct testing when a writing test is designed to require test-takers to write something. On the other hand, if a test attempts to assess test-takers’ productive skills such as speaking and writing through items that do not require them to speak or write actually, such as through multiple-choice questions, or dialogue completion tasks, it is called indirect testing.

The problem with this distinction is that even the tests alleged to be direct tests are not actually as direct as they are believed to be. One caution needs mentioning here, though. The problem is not with direct tests or tasks, but with the label they are given (Carr, 2011). This distinction between competence and performance is what constitutes the problem as it is the performance in direct testing that is scored, although performance itself is an indication of competence in truth. Therefore, familiarity with the task, the content, poor health, test anxiety, etc., may interfere with a test-taker’s performance and lead to bad performance even if they have the competence.

2.2.2.2.5 Discrete-point vs. integrated tests

One final dichotomy in the classification of test types to be mentioned in this chapter is that between discrete-point and integrated tests. As explained by Carr (2011), both of these approaches have their pros and cons; so, test designers are often faced with situations where they have to do careful thinking regarding several trade-offs when combining or choosing between these two types. If a test uses a set of separate items or tasks not connected to, or independent of each other in order to assess a distinct piece of language ability, or a trait, it is called a discrete-point test (Brown, 2004). This has traditionally been done using multiple-choice questions in standardised language tests of reading, listening, vocabulary and grammar. While this approach can be criticised for lacking authenticity, as the real-life use of language abilities and areas do not occur in isolation but in certain combinations, it provides several advantages such as more accurate or valid scoring, and satisfying the unidimensionality assumption of the IRT, a powerful statistical methodology used in the analysis of tests and test items.

On the other hand, because discrete-point tests lack authenticity, language testers have increasingly used what is called integrated tests, which intend to assess multiple aspects of language ability to simulate real-life situations. This is often done by providing test-takers with some form of language input in one, or more than one, language skill such as reading or listening, and then asking the test-taker to react to the input in another skill such as speaking or writing. This kind of approach is frequently used by language tests that set out to integrate authenticity and communicative language use into their assessment activity. Even though integrated testing is more likely to satisfy these needs, it comes with its own problems, the most prominent of which is difficulty with score interpretation. For instance, a test-taker with a high score in a task that integrates listening and speaking can be considered to be successful in both listening and speaking abilities. However, it may be difficult to exactly locate the weakness or problem in a test-taker's language ability if he or she has a low score from the same task. The problem may lie with the test-taker's listening skill, or speaking skill, or both. Therefore a good test is supposed to address this problem by having a trade-off between discrete-point and integrated tasks, usually through designing a reasonable combination of both types of tasks.

2.2.3 Qualities of a Good Test

As argued by Brown (2004), assessments of all kinds need to possess some basic qualities in order to be effective. These qualities, or principles, are practicality, reliability, validity, authenticity and washback (or test impact).

2.2.3.1 Practicality

The first principle, practicality, does not have much to do with test content directly, but is rather concerned with how efficient it is to administer a test, although decisions related to practicality issues may have profound effect on the design and planning of the test content. It addresses issues such as cost, time management, scoring and result analysis. A test can be argued to be practical as long as it is not extremely expensive for potential test-takers, not too long to manage within specified time constraints, not too difficult to administer in the field, and has a useful and time-efficient procedure for scoring and evaluation. For that reason, conditions for a test that determine its practicality may be context-dependant. For example, a test that costs $300 may be practical in the United States, but not elsewhere. Or, a test that contains 5 process-writing tasks over a semester may be practical in a classroom setting, but not in a large-scale high-stakes proficiency test for both timing and scoring difficulties.

2.2.3.2 Reliability

The term reliability is often used to refer to the scoring consistency of tests (Bachman, 1990; Carr, 2011), which can be analysed through the use of a number of statistical and mathematical methods from the point of the test administrator; however, Brown (2004) also mentions some other student-related factors contributing to reliability or lack of reliability such as fatigue, poor health and anxiety. As argued by Carr (2011), the scoring consistency of a test is usually referred to as reliability if it is a norm-referenced test, and dependability if it is a criterion-referenced test.

As reliability is related to scoring consistency, it is concerned with finding out about the sources and effect of scoring error, and these sources could be related to test methods, test-takers and also could be random. Each test is assumed to have a degree of random error, which can be minimised through systematic, well-developed and valid testing tools. If there are errors related to test methods, they can be systematic, and systematic errors could lead to test bias and inaccurate and inconsistent scoring, and thus unreliable test results. Myriad methods have been developed to address this issue, and despite the abundance of these methods, they can be divided into two groups in a broad sense based on the approach they adapt towards assessment.

The first category of these methods are those developed in line with the Classical Test Theory, or True Score Theory, which is a body of related psychometric theory that predicts the outcomes of assessment such as item difficulty and item discrimination. Because the methods used in this approach is greatly dependant on the overall ability level of the test-taker group, and the results would vary from one group to another, they are more suitable to be used with norm-referenced tests. The methods used within this approach include parallel tests, where two different tests considered to be the measures of the same ability are given to test-takers and the correlation between the two tests is calculated; internal consistency reliability analyses such as split-half reliability estimates, where a test is divided into two halves, and the correlation between them is calculated; and inter-rater and intra-rater reliability estimates.

However, due to a number weaknesses, primarily associated with group dependence, of the CTT methodology, psychometricians have come up with a number of scoring methods within Item Response Theory (IRT), which is also a body of related psychometric theory providing a foundation for scaling test-takers and items based on their responses to the items. IRT models, with the central focus on unidimensionality, meaning that each item focuses on assessing one certain latent trait or piece of ability, relate item responses to individual test-taker characteristics and item characteristics; in other words, they relate test-taker and item parameters to the probability of a discrete outcome, such as a correct response to an item; therefore, these models are group-independent. So, they are more suitable to be used with criterion-referenced tests. The models attempt to provide scoring consistency through methods such as calculating item characteristic curves (ICC), estimating ability scores, item information functions, and test information functions.

Reliability is often considered to be a related but distinct quality from validity, which will be discussed in following paragraphs. While it is true that validity is the most important quality and the ultimate objective of any assessment activity, reliability is a crucial condition for validity. Given the systematic effects of test methods, this fact applies to language assessment to a great extent, that is, the distinction between validity and reliability becomes vague due to the fact that test methods in language assessment influence both validity and reliability (Bachman, 1990).

2.2.3.3 Validity

Arguably the most important quality of a test, and the most important concept in educational and psychological assessment, validity has traditionally been defined as the "extent to which inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment" (Gronlund, 1998, p. 226). Another classical understanding of validity is the extent to which a test "measures accurately what it is intended to measure" (Hughes, 1989, p.22). Both of these definitions entail some fundamental philosophical questions regarding the very existence of a test in question, which makes the property of validity such a significant concept. However, perception of and approaches to this central concept have changed dramatically since the early days of educational and psychological assessment. Therefore, in order to gain a better insight as to how current approaches to validity work, it is necessary to have a brief look at the evolution of the concept of validity in assessment.

A summary of how the concept of validity was viewed in early assessment theory is provided in the following paragraphs based on Carr (2011), and Fulcher and Davidson (2012). In its early days, validity was roughly divided into three categories: criterion-oriented validity, content validity and construct validity. What is meant by criterion-related validity is the degree to which the test's 'criterion' has been achieved. This type of validity is often divided into two categories: concurrent validity and predictive validity. Concurrent validity is evaluated by supporting the results of a test with other performance that is concurrent beyond the test itself. For instance, the concurrent validity of a foreign language proficiency test can be supported by the actual good foreign language proficiency of a student who had a high score from the test. The second category in criterion-related validity is predictive validity. Predictive validity refers to a test's capacity to be able to predict future performance, which becomes more important in placement tests, where student aptitude - potential to learn - is highly valued.

The second category of validity in the early days of validity research was identified as content validity. Content validity refers to the extent to which a test's content is a sample that represents the domain that the assessment intends to test. For example, and academic listening test can be claimed to lack content validity if its content does not contain sufficient amount of listening input with academic content and context. According to Carroll (1980, p. 67), ensuring content validity of an EAP (English for Academic Purposes) test requires the test designers to make a description of test-takers, analysing their "communicative needs", and identifying the content of the test based on their needs. Fulcher (1999) also argues that the main challenge for early communicative language testing efforts in terms of content validity was about how to draw the best sample representing the needs of the learners and the target domain.

The third broad category of the early validity theory was construct validity. Construct validity was at the time defined by Cronbach and Meehl (1955, p. 282) as the extent to which "a test could be interpreted as a measure of some attribute or quality which is not operationally defined". In other words, a psychological construct was assumed to exist, and it needed to be operationally defined so that the assessment instrument could suggest presence or absence of this construct.

It is worth mentioning two more validity types in the early years of the validation theory before moving on with the evolution of the approaches to the concept of validity. These are consequential validity and face validity (Brown, 2004). Consequential validity is concerned with all consequences of an assessment activity, including the accurate measurement of the intended criteria, how it impacts test-takers' preparation for the test, how it affects the learning and teaching processes, and the intentional and non-intentional social consequences of the use and interpretation of a test.

Face validity, which is actually an extension to consequential validity, is related to the degree to which test-takers consider the assessment to be fair, relevant, and useful for improving learning (Gronlund, 1998). Face validity was also defined by Mousavi (2002, p. 244) as the extent to which an assessment tool "looks right", that is, appearing to be able to test the traits or constructs it aims to test, and this extent is subjectively judged by the different stakeholders of the test including test-takers, test developers and test administrators, and other “psychometrically unsophisticated observers.”

However, the fundamental philosophical assumptions of the early approaches of the validity theory began to be questioned by the logical positivists (Fulcher & Davidson, 2012), who claimed that propositions that we could not verify relative to empirical evidence did not make much sense, and thus they were not only false but also meaningless, which translated into a new assumption in the fields of psychological and educational testing and assessment that if hypotheses based on the relationship between observable variables and constructs, or between constructs, cannot be tested, then theory is not meaningful, and thus not “scientifically admissible” (p. 10).

Influenced by these philosophical enquiries, the fields of psychological testing, educational measurement and language testing have made validation studies their central focus. One of the most important contributions to this inquiry since the 1970s came from Messick (1989), who argued that evidence related to content and criterion provided information for and made contribution to score meaning, and therefore, content-related and criterion-related validity came to be recognised as two aspects of construct validity, which means that there is actually one of type of validity, which is construct validity.

Shepard noted in 1993 that although construct validity was regarded as the weaker sister to the other types of validity when it was first introduced to the study of validation, now it became much more important, fundamental even. It came to such prominence that now criterion-related and content-related validity began to be regarded as supporting evidence types to construct validity rather than being validity types on their own. She referred to construct validity as “the whole of validity theory” (1993, p. 418). This view has been made official since then by the broader field of psychological and educational measurement as well. Validity is defined by the *Standards for Educational and Psychological Testing* in relation to construct validity, calling it as “the degree to which evidence and theory support interpretation of test scores entailed by proposed uses of tests” (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999, p. 9).

This approach has given rise to what is called validity argument. According to this view, which focuses on evidence, an argument must be constructed in a way to make it possible to combine test data or other supporting information from the test in order to justify not only the inferences based on test score but also anticipated or proposed uses of the test. Carr (2011) resembles the validity argument to a court case in that it creates an explicit interpretive argument on the basis of reasonable assumptions. In the end, the argument endeavours to explore the presence or absence of construct validity, and whether the test is appropriate in terms of its stated purpose. It is argued by Chapelle (1999) and Kane (1992, 2001) that the validity argument starts with an interpretive argument and then collects and analyses evidence which supports that argument where the argument that is grounded on one or several certain score-based inferences and uses of tests. The fact that the argument is related to both score-based inferences and test uses makes it necessary that it brings together "concepts, evidence, social and personal consequences and values" (Cronbach, 1988, p. 4).

One final implication of validation arguments is related to test fairness. In fact, test fairness is a very broad area, and it involves test quality management, test administration and scoring, reasonable representation of the content to be tested, equal opportunities to learning and equal access to testing, and absence of item bias (Kunnan, 2000; Shohamy, 2001). Psychometricians, however, often focus on item bias analyses as they can be measured through a wide range of psychometric methods. An item free of item bias can be defined as an item that is able to assess the trait or ability intended to be measured without being influenced by any construct-irrelevant factors caused by any test-taker background aspects (McNamara, 2006). Favouring of one group over another based on test-taker background characteristics such as sex, age, disability, L1, socioeconomic status, or place of birth is an undesirable situation, jeopardising not only the fairness but also the validity of a test. Item bias or test bias occurs when an item or a test systematically disadvantages one group of test-takers in favour of another group when the ability level of the two groups is otherwise equal. A number of statistical and psychometric methods have been developed to identify or investigate both item bias and test bias within both CTT and IRT.

2.2.3.4 Scoring

Just as constructs are related to the "what of language testing", scoring is about "how much or how good of language testing" (Fulcher & Davidson, 2012, p. 91). Scoring helps make sense of the data or evidence collected from an assessment activity, and thus, how scoring is conducted has implications on how the performance is evaluated.

Carr (2011) states that one way of interpreting test results depending on different perspectives is based on the distinction between norm-referenced and criterion-referenced tests. The score of a test, particularly when expressed in terms of the number of correctly-answered questions, does not make much sense in its own right.

Evaluating a test score with a certain reference, or comparison, helps the evaluation process to gain meaning (Bachman, 1990). The comparison can be established either with other test-takers who took the same test, or a set of criteria determined in advance prior to testing. This difference is reflected on the distinction between norm-referenced and criterion-referenced tests. In norm-referenced testing, a test-taker's test score is construed by comparing the score against the scores of other test-takers. Therefore, test-taker scores in such tests are usually reported in percentiles. A percentile refers to any of the 99 numbered points that divide a ranked set of scores into 100 parts, each of which comprises 1/100 of the total. So, a percentile score indicates what percent of other test-takers scored equal to or lower than them on the test. Obviously, because percentile scores imply percentages, and thus, large numbers of people, they are often used in large-scale tests; or else, it would not be meaningful to split test-takers into 100 groups. However, it is still possible to use them statistically in order to make a comparison among students in the classroom. While the score comparison may be made against all the other test-takers, it could also be made against a norming sample, usually in the case of large-scale testing, that is, a group of test-takers representing the actual test-takers who took the test as part of a pre-test activity before the operational use of the test, if the number of test-takers on the test is high, consisting of tens, or hundreds, of thousands of people.

Because norm-referenced test score interpretation has the drawback of group dependence, that is, the comparison is made against other test-takers, it only provides information on test-taker success relative to the success of other test-takers, which makes it impossible to infer the scores in terms of certain learning outcomes or ability descriptors. This challenge imposed by norm-referenced score interpretation is addressed by criterion-referenced test score interpretation, where comparison is made against a set of predetermined criteria, so that the test provides scores or results that can have some absolute meaning in terms of language ability (Carr, 2011). As criterion-referenced tests are concerned with how much knowledge or ability a test-taker shows in relation to predetermined criteria, the scores are usually reported in terms of percentage, rather than percentile, which makes it possible for a test to be able to be passed by all test-takers on the condition that they pass a certain level on the test. This certain level is referred to as a cut score, that is, the minimum score for meeting the criteria defined by the test.

Bachman (1990) provides an effective and operational outline of scoring methods in language assessment. In the development of scoring procedures, as the scoring procedures make up a fundamental part of the operationalisation of the construct definition, a method must be established to allow for the quantification of the responses produced by test-takers. There are two broad categories of scoring methods. On the first category, the number of tasks accurately completed on a test is defined as the score, and thus, the number of correct responses is added. Therefore, it is necessary, in this approach, to identify a scoring method through providing a definition of the criteria as to what exactly successful completion means, and deciding whether responses will be counted as right or wrong, or with varying extents of correctness. This approach is often used with close-ended or limited response items in a test.

The other approach is more often used with test tasks that require test-takers to use productive components of language ability, such as speaking or writing tasks. In this approach, several levels on one or more rating scale of language ability are identified, and it is followed by rating of the responses to the task by raters based on these scales. The hierarchical levels on the scale are defined as an evidence of the criterion at the lowest, and as mastery at the highest. Such scales also enable test administrators to provide test-takers with meaningful feedback on their abilities. One important caveat worth noting is that as with any other decision in language test development, scoring decisions are to affect and be affected, and thus be in compliance, with all the other assessment decisions.

2.3 Assessment and Teaching

Assessment has an undeniable role in teaching and learning processes. According to White (2009), learning is initiated by assessment, which acts as an engine. It is indeed wrong to treat teaching and assessment as separate constructs, for the processes of teaching and learning involves assessment as an inherent component, and teachers allocate a great deal of their professional time to assessment and activities related to assessment. The quality of the instruction and learning depends upon the quality of the assessment tools being utilised, which makes the need for good assessment practices crucial (Stiggins, 1999). Using these good assessment practices and tools that can serve as good informants, teachers can obtain a number of benefits such as adjusting the pace of the lesson, coming up with decisions about whether the content of the course is relevant or irrelevant, and whether the instruction is effective or ineffective, as well as helping learners build up confidence for standardised tests.

The profession of language teaching leaves the language teacher with the responsibility of assessment (Mertler, 2003). The concepts of teaching and assessment affect one another, and they are informed and improved by each other (Malone, 2013). For this reason, teachers are supposed to have a role in bridging between the concepts of teaching and assessment. Therefore, their salient role in assessment processes has been acknowledged by Stiggins, (1999) and Popham (2009) who concluded that teachers can make more informed decisions once they are equipped with knowledge of assessment. Teachers' crucial role in assessment means that their assessment knowledge has a great impact on the quality of instruction and on education in general (Malone, 2013).

According to Price et al. (2012), language teachers are expected to be knowledgeable enough in processes related to assessment. Yet, research findings suggest that language teachers do not think that they are adequately equipped with assessment knowledge (Plake, 1993). Stiggins (2010) stressed the seriousness of the problem with an interesting quotation: "assessment illiteracy abounds" (p. 233), which suggests that whether teachers are adequately equipped with LAL knowledge is open to controversy despite the role expected of them.

2.4 Assessment Literacy

This section examines AL by discussing how the concept has been defined and the reason why it is has been deemed important by educational researchers.

2.4.1 Definition of AL

AL can be considered to be a basic understanding of educational assessment, and skills related to it (Stiggins, 1991). According to Wiggins (1998), AL involves techniques and concepts that educators should have a knowledge of while designing and using assessment tools. He adds that what is learned by learners and to what extent they meaningfully engage in with what is learned are affected by the nature of assessment.

AL goes beyond simply having knowledge of test formats like constructed-response items, multiple-choice items, cloze tests, matching activities, etc. Fundamentally, it covers having mastery of assessment principles (McMillan & Nash, 2000). It is about making assessment-related decisions regarding what assessment tools to use, why and how to use them. According to a number of researchers (Calfee & Masuda, 1997; McMillan, 2001; Sanders & Vogel, 1993; Stiggins & Conklin, 1992), educators must have knowledge of essential assessment principles, concepts, techniques and procedures in order to make sound and safe decisions.

According to Stiggins (1995), teachers who are assessment literate are aware of

> … 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 … (p. 240)

This quote suggests that AL refers to the knowledge of assessment as well as the application of this knowledge to the practices of assessment.

However, there is no consensus on how best to define AL. The majority of the definitions formulated so far are either context-bound or imply a specific content area. Xu and Brown (2016) provide an extensive overview of the definitions of AL. The method most widely used in making a definition of AL makes use of specific knowledge, understanding of and skills related to assessment that an educator who is assessment literate must have (Boyles, 2005; Gareis & Grant, 2015; Popham, 2004; Stiggins, 1991; Xu & Brown, 2016).

Popham (2009) proposed a widely-cited and popular definition of AL through a list of content points. This list emerged out of an existing version of AL standards developed by a team of field professionals at the Michigan Assessment Consortium. The list developed by Popham specifies the content areas that are needed by educators to be assessment literate. Table 1 presents the list of criteria (Popham, 2009, pp. 8-10). According to the author, these content points are to be gained by teachers through training and professional development.

Table 1: Popham's (2009) Suggested Content Points for Teacher AL

Point	Explanation
1	The fundamental function of educational assessment, namely, the collection of evidence from which inferences can be made about students' skills, knowledge, and affect
2	Reliability of educational assessments, especially the three forms in which consistency evidence is reported for groups of test-takers (stability, alternate-form, and internal consistency) and how to gauge consistency of assessment for individual test-takers
3	The prominent role three types of validity evidence should play in the building of arguments to support the accuracy of test-based interpretations about students, namely, content-related, criterion related, and construct-related evidence
4	How to identify and eliminate assessment bias that offends or unfairly penalizes test takers because of personal characteristics such as race, gender, or socioeconomic status
5	Construction and improvement of selected response and constructed-response test items
6	Scoring of students' responses to constructed-response tests items, especially the distinctive contribution made by well-formed rubrics
7	Development and scoring of performance assessments, portfolio assessments, exhibitions, peer assessments, and self-assessments
8	Designing and implementing formative assessment procedures consonant with both research evidence and experience-based insights regarding such procedures' likely success

Table 1 (continued)

9	How to collect and interpret evidence of students’ attitudes, interests, and values
10	Interpreting students’ performances on large-scale, standardized achievement and aptitude assessments
11	Assessing English Language Learners and students with disabilities
12	How to appropriately (and not inappropriately) prepare students for high-stakes tests
13	How to determine the appropriateness of an accountability test for use in evaluating the quality of instruction

The definition proposed for AL by Popham (2009) suggests that assessment literacy refers to the understanding of the fundamental concepts and procedures in assessment that are likely to have an influence on decisions to be made in the classroom (classroom assessment) as well as decisions made inside and outside of the classroom (accountability assessment).

To be more precise, AL refers to the understanding of both assessment concepts and contextual procedures that influence the decision-making process. Understanding of basic assessment concepts is similar to the definition provided by Xu and Brown (2016) in that it emphasises the importance of having the knowledge of the terminology and concepts of assessment. However, understanding of contextual procedures influencing the decision-making process implies translation of the knowledge of assessment into practice that would impact educational outcomes. Therefore, AL involves how a teacher selects, employs and interacts with assessment both inside and outside of the classroom.

Apart from the list and definition of skills regarding AL provided by Popham (2009) and the Michigan Assessment Consortium, there are several other such lists developed by other researchers and institutions. Another example of such a list, of utmost importance to the present study, is a document named the *Standards for Teacher Competence in Educational Assessment of Students* (AFT, NCME, & NEA, 1990). The document offers seven standards of teacher AL (Table 2).

Table 2: Standards for Teacher Competence in Educational Assessment

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

The *Standards* document was created in order to emphasise the importance of classroom assessment. These standards address teachers' classroom-based assessment competencies as well as the role played by teachers in the decision-making process beyond the classroom. They depart from teachers' role in the classroom and progress toward their role within the broader educational community. The details of each standard are presented below (retrieved from: https://buros.org/standards-teacher-competence-educational-assessment-students):

1. **Standard 1: Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.** Teachers who meet this standard will have the conceptual and application skills that follow. They will be able to use the concepts of assessment error and validity when developing or selecting their approaches to classroom assessment of students. They will understand how valid assessment data can support instructional activities such as providing appropriate feedback to students, diagnosing group and individual learning needs, planning for individualized educational programs,

motivating students, and evaluating instructional procedures. They will understand how invalid information can affect instructional decisions about students. They will also be able to use and evaluate assessment options available to them, considering among other things, the cultural, social, economic, and language backgrounds of students. They will be aware that different assessment approaches can be incompatible with certain instructional goals and may impact quite differently on their teaching.

2. **Standard 2: Teachers should be skilled in developing assessment methods appropriate for instructional decisions.** Teachers who meet this standard will have the conceptual and application skills that follow. Teachers will be skilled in planning the collection of information that facilitates the decisions they will make. They will know and follow appropriate principles for developing and using assessment methods in their teaching, avoiding common pitfalls in student assessment. Such techniques may include several of the options listed at the end of the first standard. The teacher will select the techniques which are appropriate to the intent of the teacher's instruction.
3. **Standard 3: The teacher should be skilled in administering, scoring and interpreting the results of both externally-produced and teacher-produced assessment methods.** Teachers who meet this standard will have the conceptual and application skills that follow. They will be skilled in interpreting informal and formal teacher-produced assessment results, including pupils' performances in class and on homework assignments. Teachers will be able to use guides for scoring essay questions and projects, stencils for scoring response-choice questions, and scales for rating performance assessments**.** They will be able to use these in ways that produce consistent results.
4. **Standard 4: Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.** Teachers who meet this standard will have the conceptual and application skills that follow. They will be able to use accumulated assessment information to organize a sound instructional plan for facilitating students' educational development. When using assessment results to plan and/or evaluate instruction and curriculum,

teachers will interpret the results correctly and avoid common misinterpretations, such as basing decisions on scores that lack curriculum validity. They will be informed about the results of local, regional, state, and national assessments and about their appropriate use for pupil, classroom, school, district, state, and national educational improvement.

5. **Standard 5: Teachers should be skilled in developing valid pupil grading procedures which use pupil assessments.** Teachers who meet this standard will have the conceptual and application skills that follow. They will be able to devise, implement, and explain a procedure for developing grades composed of marks from various assignments, projects, in class activities, quizzes, tests, and/or other assessments that they may use. Teachers will understand and be able to articulate why the grades they assign are rational, justified, and fair, acknowledging that such grades reflect their preferences and judgments. Teachers will be able to recognize and to avoid faulty grading procedures such as using grades as punishment. They will be able to evaluate and to modify their grading procedures in order to improve the validity of the interpretations made from them about students' attainments.
6. **Standard 6: Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.** Teachers who meet this standard will have the conceptual and application skills that follow. Teachers will understand and be able to give appropriate explanations of how the interpretation of student assessments must be moderated by the student's socio-economic, cultural, language, and other background factors. Teachers will be able to explain that assessment results do not imply that such background factors limit a student's ultimate educational development. They will be able to communicate to students and to their parents or guardians how they may assess the student's educational progress. Teachers will understand and be able to explain the importance of taking measurement errors into account when using assessments to make decisions about individual students. Teachers will be able to explain the limitations of different informal and formal assessment methods. They will

be able to explain printed reports of the results of pupil assessments at the classroom, school district, state, and national levels.

7. **Standard 7: Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.** Teachers who meet this standard will have the conceptual and application skills that follow. They will know those laws and case decisions which affect their classroom, school district, and state assessment practices. Teachers will be aware that various assessment procedures can be misused or overused resulting in harmful consequences such as embarrassing students, violating a student's right to confidentiality, and inappropriately using students' standardized achievement test scores to measure teaching effectiveness.

The standards focus on assessment knowledge and skills needed by teachers in relation to the activities and practices (a) before instruction, (b) during instruction, (c) after instruction, (d) decision-making processes in the school context, and (e) decision-making processes within the context of the educational community. The following paragraphs present other examples of frameworks of AL.

2.4.1.1 Frameworks for AL

A number of frameworks for AL have been developed based on a variety of definitions and manifestations of the concept, and each outlines its structure. Such frameworks have frequently focussed on the practices of classroom assessment and generally been intended for classroom teachers. The *Standards* document developed by AFT, NCME, and NEA (1990) is an example of this paradigm. These frameworks usually aim to bridge the AL gaps experienced by pre-service and in-service teachers. Siegel and Wissehr (2011) developed one such framework concentrating on AL of pre-service teachers. The framework addresses classroom assessment principles as well as teachers’ knowledge of assessment procedures and assessment instruments.

Another framework focussing on classroom teachers was developed by Gareis and Grant (2015). The framework divides AL into three domains for teachers and administrators: (a) types of measures, (b) quality of measures, and (c) results and their uses. Kahl, Hofman, and Bryant (2013) have lately proposed the Assessment Literacy Domain that draws on the existing standards generated by a variety of institutions. They advise that the framework used for assessment should focus on benefiting from the standards in a way to make use of assessment results to inform instruction, programs and assessment design. The teacher is placed by these frameworks at the centre of assessment, and expected to use a number of abilities related to assessment, measurement and interpretation of assessment results.

Some AL frameworks set out to include elements of professional development. This means going beyond a mechanical list of skills needed by teachers to be assessment literate. The motive of such frameworks emerges from the idea that many teachers are not trained with particular focus on assessment, and so they should be aided to possess a certain degree of assessment knowledge (DeLuca & Klinger, 2010; Popham, 2009; Volante & Fazio, 2007; Wang, Wang, & Huang, 2008). Similarly, the social-constructivist AL perspective of Inbar-Lourie (2008) highlights the importance of developing AL through professional development. This need was particularly underlined for language teachers. The framework proposed by Inbar-Lourie (2008) suggests that how AL is defined depends on the context or content. In other words, assessment literacy needed by language teachers would be influential in how the training for the development of assessment literacy should be designed, which shows that AL is context-bound, and thus a separate set of assessment knowledge may be needed by each teacher.

Xu and Brown (2016) also proposed a framework focussing on teachers and highlighting the importance of creating a pathway for developing AL that covers all of the stages involved in teacher education and professional development of teachers. This framework of AL, called Teacher Assessment Literacy in Practice (TALiP), involves five components: (a) teacher conceptions of assessment, (b) institutional and socio-cultural contexts, (c) TALiP, the core of the framework, (d) teacher learning and (e) teacher identity, (re)construction as assessors. These components reflect the three assessment knowledge domains that a teacher must possess in order to become assessment literate; i.e., (a) educational assessment knowledge, (b) knowledge of the interconnectedness of assessment, teaching, and learning, and (c) the assessor identity. The first two of these domains reflect the domains proposed by the previously mentioned frameworks, but assessor identity emphasises the role and context of the teacher in a similar way to the framework of Inbar-Lourie (2008).

The frameworks mentioned so far emphasise the traditional components of classroom assessment. Nevertheless, other concepts or components such as data and measurement are important. These frameworks address the concepts of measurement together with other components. Popham was one of the first researchers to highlight measurement as a distinct concept in classroom assessment literacy, at a time when the concepts of assessment, measurement and testing were widely used interchangeably (Daniel & King, 1998; Popham, 1995).

An empirical example of such an approach combining measurement and assessment came from Daniel and King in 1998, where the researchers asked teachers about how familiar they were with the basics of measurement such as content validity, reliability, correlation range, standard error of measurement, mean, mod, and median. The teachers were even asked to form applied judgements about the concepts (such as interpretation of correlation coefficients). Such studies propose that measurement cannot be considered separate from classroom assessment, and thus should be a central part of teacher classroom AL.

A "working" (not necessarily technical) knowledge of measurement principles could benefit teachers, as suggested by Brookhart (2001), who studied this issue by making sure that the questions attempted to test teacher AL although being based on measurement knowledge. Teachers' need for practical knowledge was emphasised by this research study. Such studies highlight the need for the incorporation of measurement to assessment training in teacher training programs so that teachers have an adequate level of understanding of the concepts related to measurement as they encounter them in the field.

On the other hand, inventories or tests of AL such as Assessment Literacy Inventory (ALI; Mertler & Campbell, 2005) and the Assessment Knowledge Test (AKT; Wang et al., 2008), focussing on teachers, openly test teachers' knowledge of measurement concepts such as percentile, reliability, and cut score. They stress how important it is to teach basic measurement principles to teachers because a teacher's overall AL benefits from them (Mertler & Campbell, 2005; Wang et al., 2008). According to Brookhart (2001), a theoretical or technical understanding of a concept such as the standard error of measurement is not crucial for a teacher, but an intersection exists between the suggested assessment knowledge (Brookhart, 2011; Popham, 2011) and the recent assessment literacy measures and inventories.

Measurement knowledge is included in the recent measures of AL. Therefore, it may be a good idea to achieve a balance, and to incorporate the basic principles of measurement in the teachers' AL.

There are also other frameworks that put specific emphasis on data and professional development in the development of teacher AL. This approach is referred to as data literacy. Data literacy proposes that a teacher who is assessment literate is knowledgeable about when and how to apply assessment skills and assessment knowledge within a certain context. The framework developed by Supovitz (2010) for data-related professionals has four stages, which are (a) data capturing, (b) meaning making, (c) information sharing, and (d) knowledge codification. Teachers play a key role in this framework, as they are the actors who capture data. This approach has been further developed by Jimerson and Wayman (2015) who suggested that individual learning and organizational learning are supported by one another. That is to say, both group learning and individual learning are important for effective data-related learning. It is suggested that knowledge of assessment gained by a teacher in the classroom plays an important role in the broader educational community through communities of practice as he or she collaborates with other teachers in learning how to manage data.

As is the case for the AL frameworks discussed in the previous paragraphs, data literacy frameworks have also had teachers at their centre. A data literacy framework developed by Gummer and Mandinach (2015), for instance, focusses on three domains: (a) disciplinary content knowledge, (b) pedagogical content knowledge, and (c) data use for teaching knowledge and skills. Content knowledge, classroom practices and pedagogy are some of the domains present also in many other frameworks; yet, data plays a much more important role within this framework.

Last but not the least, a systematicity framework was proposed by Athanasas, Bennett, and Wahleithner in 2013, in which data literacy is informed. According to the researchers, data collection, data analysis and information use for teaching are incorporated by data literacy. This framework is similar to many of the frameworks mentioned so far in that they treat the evidence collected as data.

Supported by the results of a substantial amount of research, these frameworks emphasise the need for a definition of the AL construct to be agreed upon by everyone. However, no unified definition of AL has been reached; nor has there been any framework proposed that covers the full range of aspects or components existing within AL. One reason for this may be the fact that the structure and nature of AL is context-bound, as stated by Inbar-Lourie (2008). An AL framework needed by a teacher in one context may not be needed by another teacher in another context.

2.4.1.2 Measurement of Assessment Literacy

Even though a complete evaluation of a teacher's understanding of assessment literacy would require to observe their assessment practices and decisions made both inside and outside the classroom in addition to testing their AL at the knowledge base, most of the objective measures developed to investigate teachers' assessment competencies and skills have focussed on the knowledge base due to practicality reasons, and as the knowledge base is considered to be the first underpinning of AL and it is considered to be a fundamental component contributing to the success and effectiveness of the implementation of assessment (Xu & Brown, 2017).

A number of self-reported measures have been designed to elicit information from teachers regarding their AL, and most of these studies have focussed on teachers' strengths and weaknesses to identify their training needs (Fulcher, 2012; Hasselgreen, Carlsen, & Helness, 2004; López & Bernal, 2009). On the other hand, as research has shown that self-reported measures or self-evaluation could lead to inaccurate information, there has been a tendency to develop and administer more objective assessment instruments to acquire relatively more reliable data from teachers regarding their AL, and most of these instruments use multiple-choice questions. Generally, the measures have been developed or adapted based on the frameworks of AL discussed in the previous section. A list of studies that have used some of the objective measures to directly test the assessment knowledge of teachers are presented in Table 3.

Some of the measures included in Table 3 developed their own multiple-choice questions while some adopted or adapted questions from other measures. The Criterion-referenced Assessment Literacy Test shown in the table is a measure aiming to test teachers' AL focussing on criterion-referenced and norm-referenced tests, the concepts of validity and reliability, and misuse of assessment data. The

Measurement Knowledge Test focusses on norm-referenced tests, use of standardised scores, and proficiency levels. The Teacher Assessment Literacy Questionnaire (TALQ), Assessment Literacy Inventory (ALI) and Classroom Assessment Literacy Inventory (CALI) are all based on the *Standards* document.

In spite of producing differing results regarding internal consistency reliabilities, all of the studies mentioned have reported limited teacher assessment literacy for performing high quality assessment.

2.4.2 Significance of AL

Researchers increasingly regard AL as central to a teacher’s teaching skills (Popham, 2009; Xu & Brown, 2016). The justification behind the increasing recognition of AL as essential to the teaching profession is that a sound mastery of the principles and techniques of assessment helps teachers arrive at sophisticated and informed decisions about the validity of assessment practices as well as educational policies (Kane, 2006; Messick, 1989). A teacher who is knowledgeable in AL can not only make accurate inferences about student learning, but also inform students and other educational stakeholders about those inferences, in turn being able to adjust instruction accordingly. On the other hand, a teacher without sufficient knowledge and mastery of AL may end up with reduced validity and reliability, and thus making erroneous judgments and ill-informed educational decisions.

Table 3: Summary of Studies Using Some of the Objective Measures to Directly Test Teacher Assessment Knowledge

Study	Measure	Number of items	Item Type	Participants	Number of Participants	Reliability Estimate
King (2010)	Criterion-referenced Assessment Literacy	24	Multiple-choice questions	Teachers and administrators	352 teachers + 28 Administrators	.73
Gotch & French (2013)	Measurement Knowledge Test	20	Multiple-choice questions	In-service teachers	650	.47
Plake (1993)	Teacher Assessment Literacy Questionnaire (TALQ)	21	Multiple-choice questions	In-service teachers	555	.54
Quilter & Gallini (2000)	TALQ	21	Multiple-choice question	In-service teachers	117	.50
Chapman (2008)	TALQ	16	Multiple-choice question	In-service teachers	61	.54
Alkharusi et al. (2012)	TALQ	32	Multiple-choice question	In-service teachers	165	.62
Mertler (2003)	Classroom Assessment Literacy Inventory (CALI)	35	Multiple-choice question	In-service and pre-service teachers	197 in-service + 67 pre-service teachers	.57 & .74
Mertler (2005)	CALI	35	Multiple-choice question	In-service and pre-service teachers	101 in-service + 67 pre-service teachers	.44 & .74
Alkharusi et al. (2011)	TALQ	35	Multiple-choice question	In-service and pre-service teachers	233 in-service + 279 pre-service teachers	.78 & .78
Mertler & Campbell (2005)	Assessment Literacy Inventory (ALI)	35	Multiple-choice question	Pre-service teachers	249	.74
Davidheiser (2013)	ALI	35	Multiple-choice question	In-service teacher	102	.82
Ryan (2018)	CALI	35	Multiple-choice question	Pre-service teachers	165	.92

Although there are numerous arguments supporting how beneficial AL can be for teachers (Brookhart, 2011), it is reported that many teachers constantly end up having to make assessment-related decisions without adequate assessment training (DeLuca & Bellara, 2013; Schafer & Lizzitz 1987). Teachers may allocate around a half to a third of their professional time on activities related to assessment (Stiggins, 1995); yet, their AL knowledge is not adequate at all (DeLuca & Klinger, 2010; Popham, 2009). Although professionals in the education sector are expected to conduct assessment in order to come up with educational decisions, there is evidence suggesting that they reach these decisions without a complete or sound understanding of educational assessment (Popham, 2006).

2.5 Language Assessment Literacy

LAL is discussed in this section through a discussion of how it is defined; how distinct it is from AL in particular, and an overview of notable research studies with findings on EFL teachers' LAL.

2.5.1 Definition of LAL

Possibly overlapping with AL, or even considered to be subordinate to it (Taylor, 2013), LAL has multiple layers and stages within it (Pill & Harding, 2013; Taylor, 2013). The stages range from a basic understanding of knowledge of measurement and assessment 'know-how' in terms of classroom practice to a better command of 'having the capacity to ask and answer critical questions about the purposes of assessment, about the fitness of the tool being used, about testing conditions and about what is going to happen on the basis of the results' (Inbar-Lourie, 2008, p. 389). However, the question of what specific expertise is required in LAL as opposed to AL has remained pertinent (Inbar-Lourie, 2013).

The use of the term 'language assessment literacy', having emerged with reference to AL, is relatively recent, yet it evokes an area distinct from AL. Various definitions of LAL have been put forward in the literature. According to Malone (2013), LAL refers to "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" (p. 329). Inbar-Lourie (2008) argued that "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" (pp. 389-390). Inbar-Lourie (2017) also noted that although the term LAL stems from AL, it implies a different meaning, in that LAL intends to “set itself apart as a knowledge base that incorporates unique aspects inherent in theorizing and assessing language-related performance (p. 259). According to Lam (2015), LAL refers to

> … teachers’ understanding 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)

Taylor (2009) also defined LAL as “the level of knowledge, skills, and understanding of assessment principles and practices that is increasingly required by other test stakeholder groups, depending on their needs and context” (p. 24). These definitions imply that it takes additional skills about language educators to acquire LAL, compared to AL. On the other hand, despite the presence of many definitions of LAL, some of which have been documented here, it would not be wrong to state that LAL “is still in its infancy” (Fulcher, 2012, p. 117).

2.5.2 Findings on EFL Teachers’ LAL

Central to AL is assessment knowledge, according to Xu and Brown (2017) who noted that assessment literacy needs to start with the investigation of its knowledge base. Several studies have addressed the current levels of EFL teachers’ LAL in various contexts. For example, Lam (2015) conducted a research study aiming to learn about how the LAL of pre-service teachers in five institutions in Hong Kong was facilitated or inhibited by two courses on language assessment. The study analysed the data gathered from the institutions about the courses and found that there was no sufficient support for LAL in the programmes. Similarly, a study conducted by Tsagari and Vogt (2017) showed that the participants, who were teachers from Cyprus (n=16), Greece (n=22) and Germany (n=25) without any assessment training, thought that they felt inadequate in terms of LAL. Also, another study done by Volante and Fazio (2007) with 69 pre-service teachers found that the self-rating participants had very low LAL scores. It was found that the participants used assessment and assessment tools primarily for traditional summative purposes.

A few studies so far have looked at EFL teachers’ LAL in Turkish context. A study conducted by Hatipoğlu (2015) with 124 pre-service teachers, aiming to find out about the assessment knowledge of pre-service teachers as well as what they expected from their testing course, found that the four-year ELT programme did not equip them with adequate assessment knowledge, and that their expectation from the course was to help them evaluate and select learners, and write tests in addition to helping them prepare their students for exams. Öz and Atay (2017) also carried out a study to investigate Turkish EFL teachers' perceptions of classroom language assessment and how it is reflected in their classroom activities. The findings of the study demonstrated that even though they know about the basics of classroom assessment, they had difficulty in translating this knowledge base into practice. Another study by Mede and Atay (2017) used an assessment literacy scale adapted from Vogt and Tsagari (2014). The study had 350 participants who were EFL teachers and found that the teachers did not have adequate levels of LAL, and were in need of training in many subjects related to assessment. Finally, Ölmezer-Öztürk and Aydın (2018), conducted a study with 542 participants (ELF teachers) from 53 universities (37 state universities and 16 private universities) in Turkey. They came up with similar results and found that EFL teachers had limited knowledge in assessment-related issues, although they scored higher in assessing reading compared to the assessment of the other skills. They found that the only sub-group who had higher LAL levels were EFL teachers who were members of a testing unit, which suggests that hands-on practice helps teachers improve their LAL.

CHAPTER 3

METHODOLOGY

3.1 Purpose

Teachers spend a great deal of their time on assessment-related activities, whether it is a formal or informal assessment, or whether it is a formative or a summative assessment. They may engage in such activities at different levels including classroom, local and national levels. Their assessment task goes beyond creating or developing and administering the assessments. They are also tasked with explaining the results of the assessments to learners and other stakeholders including school administrators, other teachers and parents (Kahl, Hofman & Bryant, 2013). Therefore, teachers should possess certain degrees of assessment knowledge to be able to better undertake such a responsibility.

Assessment practices have been evolving and changing rapidly, accompanied by changes in how it is carried out and reported. The increasing use of student growth percentile rather than ranking methods and other gain score is an example of this change (Betebenner, 2009; Walsh & Isenber, 2015). EFL teachers may find it challenging to have a full understanding of the link between statistics and measurement-related changes when trying to make sense of concepts such as student growth. This is an example of what Popham (2011) defined as accountability assessment. Teachers need a functional knowledge of statistics and assessment-related topics in order to grasp the relationship between instruction and assessment.

Furthermore, some other concepts and procedures related to assessment and AL have an impact on teaching and quality of teaching. Among such concepts are item and test creation, adaptation and development, administering tests, reporting test scores and evaluating the outputs of tests. The many facets of AL and its relation to and impact on teaching has spurred researchers to study both how to assess and improve teachers' AL (Popham, 2009; Taylor, 2009; Wang et al., 2008). Mertler reported in 2003 that in-service teachers across all grades and content areas admit having problems with understanding, applying and interpreting assessment-related practices; therefore, the problems associated with AL encountered by teachers could be addressed considering higher education and teacher training as a point of departure, which is the reason why the present study has chosen pre-service EFL teachers who have recently taken an English Language Assessment course as the focus of study in relation to finding out about their strengths and weaknesses with respect to LAL.

3.2 Research Questions

The aim of the current study is to explore the answers to the following research questions:

1. What are the psychometric properties of the adapted Classroom Assessment Knowledge instrument (Tao, 2014), devised to assess EFL teachers' language assessment literacy knowledge base?
2. What is the language assessment literacy knowledge base level of pre-service EFL teachers in the higher education context in Turkey?
3. What factors, if any, affect language assessment literacy of pre-service EFL teachers in the higher education context in Turkey?

3.3 Context and Participants

The current study was carried out with the participation of 4th grade students from the Department of Foreign Language Education at the Middle East Technical University (METU) and the Division of English Language Education at Gazi University, two prominent state universities in Ankara offering an English Language Teacher training program.

The primary research objective of the present study was to investigate the psychometric properties of a measure developed with the purpose of assessing EFL teachers' AL. However, a second research objective was to find out about the current knowledge of pre-service EFL teachers' in the higher education context in Turkey. 4th grade pre-service EFL teachers were chosen as participants in the study because they are believed to represent a group of stakeholders in the foreign language education sector in Turkey that would soon need such knowledge in their career. Also, they are assumed to be among the stakeholders that represent one of the most knowledgeable groups in terms of assessment topics as they have recently taken courses on assessment and language assessment. Therefore, having recently been taught on the subjects of assessment and language assessment, they were expected to demonstrate a substantially good performance on the measure that assesses their knowledge on the assessment topics.

The Department of Foreign Language Education at METU offers B.A., M.A., and Ph.D. programs in English Language Teaching. The program aims to equip prospective ELT teachers with skills necessary to understand and cope with theoretical and methodological issues in ELT. The English Language Teacher Training program at Gazi University also aims to provide through its B.A., M.A., and Ph.D. programs its students with all skills needed for the English Language Teaching profession. A total of 74 4th grade students (58 from Gazi University and 16 from METU) participated in the study. All of the students participating in the study have previously taken at least one course on English Language Testing.

Participants from METU have taken a compulsory course titled English Language Testing and Evaluation. The course used to be offered in the 7th semester before the 2018-2019 academic year, and it is now offered in the 6th semester. The objectives and learning outcomes of the course are described as follows (METU, n.d.):

> **Course Objectives:**
> At the end of this course students will
> **(**1) learn and use basic terms and concepts related to language testing appropriately where/when necessary appropriately
> (2) engage in various processes and practices related to assessment of language proficiency successfully
> (3) perform statistical analysis of testing data
> (4) design, implement and evaluate a variety of testing instruments for a specific group of language learners
> (5) acquire skills necessary for evaluating various language tests and test results/items
> **Course Content:**
> 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; statistical calculations.
> **Course Learning Outcomes**
> On successful completion of this course, students will be able to
> (1) use basic terms and concepts related to language testing appropriately where/when necessary

(2) express successfully their knowledge related to the role of tests within the curriculum design for language teaching
(3) discuss the importance of test selection according to the profile of the learners and the teaching context
(4) select tests according to the profile of the learners and the teaching context
(5) write, implement and evaluate a variety of testing instruments for a specific group of language learners
(6) use different techniques for adapting language test
(7) use various processes and practices related to the assessment of language proficiency successfully
(8) perform statistical analysis of testing data
(9) evaluate tests and test results/items
(10) discuss the advantages and disadvantages of using published and class teacher-written tests

Participants from Gazi University had previously taken a compulsory course titled Assessment and Evaluation in Education, offered in the 6th semester. The program also offers in the 8th semester another compulsory course titled English Language Testing and Evaluation. The contents of the course are listed as follows (Gazi University Department of English Language Teaching, n.d.):

Course Contents (Assessment and Evaluation in Education)
(1) The place and importance of assessment and evaluation in education
(2) Fundamentals of assessment and evaluation in education
(3) Psychometric properties of assessment and evaluation instruments (validity, reliability and practicality)
(4) Developing and administering achievement tests
(5) Interpreting test results and giving feedback
(6) Analysing tests and items
(7) Evaluation and scoring
Course Contents (English Language Testing and Evaluation)
(1) Types of tests and assessment methods for different age groups and language levels in the teaching of language skills
(2) Principles guiding the assessment and evaluation of language skills
(3) Item types used in the assessment of reading, writing, listening, speaking, vocabulary and grammar
(5) Hands-on practice for creating different item types and on test evaluation

Participation in this study was on a voluntary basis. The researcher invited the participants to the study based on convenience sampling, which Mackey and Gas (2005) define as selecting the participants that are suitable for the study (p. 222). One of the most frequently employed sampling methods in educational research; this method affords the researcher effectiveness with respect to time, money and effort (Mujis, 2004).

The participants, who were all 4th grade ELT students at METU or Gazi University, were invited to take part in the study on two separate sessions, one at METU, and the other at Gazi University. Because the scale the participants were required to take tested for knowledge, and the participants were required to allocate at least half an hour for the proctored session, the turn-up rate was relatively low, especially at METU, even though all 4th grade students at both schools were reached out and invited to participate in the study. In an effort to increase the turn-up rate and motivate the participants to demonstrate their full concentration and assessment knowledge, the participants were informed that the top three scoring participants from each school would be rewarded with a gift card each from a well-known nationwide bookstore. Both of the sessions were conducted in the spring semester of 2019-2020 academic year. A total of 74 participants from the two schools took part in the study.

As shown in the table regarding participant descriptive statistics (Table 4), 16 (21.6%) were from METU and 58 (78.4%) were from Gazi University. 59 (79.7%) of the participants were female, 13 (17.6%) were male students while 2 participants (2.7%) preferred not to specify gender. Only 7 (9.5%) of the participants had previously taken a workshop, seminar or webinar that was specifically dedicated to language assessment apart from a curriculum-based course on language assessment.

Table 4: Participant Descriptive Statistics (1)

	Number		**Percentage**	
Gender	Female	Male	Female	Male
	59	13	79.7	17.6
Previous attendance to a workshop	Yes	No	Yes	No
	7	67	9.5	90.5

The mean age of the participants was 22.2, and the participants had an average CGPA of 3.2 out of 4.00. In relation to how they perceive their preparedness level for the overall job of being a classroom teacher and for assessing student performance, the mean score for the overall job preparedness was 3.0, corresponding to ‘prepared’, and the mean score for preparedness for assessing student performance was 2.7, corresponding to somewhere between 'somewhat prepared' and 'prepared', meaning that even though the majority of the participants considered that they felt prepared for being an English language teacher, they self-reported having problems and hesitations regarding the assessment component of the teaching profession (refer to Table 5).

Table 5: Participant Descriptive Statistics (2)

	Mean	Median	Std. Dev.	Minimum	Maximum
Age	22.2	22	1.3	20	28
CGPA	3.2	3.3	0.4	2.2	4.0
Preparedness for the overall job	3.0	3.0	0.7	1.0	4.0
Preparedness for assessment	2.7	3.0	0.8	1.0	4.0

3.4 Data Collection Procedures

Before the data collection process began, the researcher obtained the permission letter (Appendix A) from the Human Research Ethics Committee at the Middle East Technical University (METU). All participants from METU and Gazi University participated in the study and completed the measure used in the study to elicit information about their assessment knowledge.

3.4.1 Data Collection Instrument

Knowledge base of AL can be considered to be the foundation of AL, as it has been acknowledged by several researchers that teachers' knowledge base greatly influences the effectiveness and success of the implementation of assessment (Bandura, 1997; Fishbein & Ajzen, 2010). As it has been widely reported, instruction and learning is considerably enhanced when teachers are assessment literate (Boud, 2006; Earl, 2013; Joughin, 2009; Tang, 1994), which suggests that teachers must possess adequate AL to help them engage in high quality assessments. Even though an educator's complete AL consists not only of the knowledge base, but also of the ability to apply the knowledge into practice, assessment researchers (e.g., Popham, 2006, 2009; Stiggins, 1991, 1995) report that knowledge base could be a good indicator of the practical success. The researchers argue that the greater a teacher's knowledge base of AL is, the more successful the teacher is in terms of the implementation of high quality assessments. Therefore, due to practicality reasons, the present study aimed to elicit information regarding AL knowledge base of pre-service EFL teachers, which the researcher believes could give some insight into the bigger picture of AL of pre-service EFL teachers at the two state universities in Turkey. The researcher made use of a modified version of an assessment instrument called CAK, adapted by Tao (2014) for a study that aimed to assess EFL teachers' assessment knowledge base.

The instrument employed by the current study was designed originally as a measure of assessment knowledge with regards to topics and concepts in the broader field of educational assessment, including but not limited to the principles of educational assessment such as validity and reliability, types of assessment, scoring and grading, interpreting assessment results, and measurement concepts such as percentile and standard deviation. Yet, it is often argued that language assessment is distinct from the broader concept of educational assessment, particularly because it contains elements specifically related to the theories of language and language development (Inbar-Lourie, 2008) such as teaching and learning of language skills and language areas. However, to the knowledge of the researcher, there is no language-specific objective measure of LAL available in the literature so far that has been widely tested and argued to generate evidence-based validity and reliability arguments. In addition, one could argue that the language-specific elements of language assessment are language education-specific reflections of the central concept of validity in the educational assessment. Therefore, the researcher decided to use an adapted and modified version of a measure of AL that has been widely used so far in a number of studies about which a substantial amount of psychometric and statistical information has been reported. In other words, although the CAK instrument does not explicitly contain language-specific assessment concepts such as the assessment of language skills and language areas, it is still likely to elicit information from the participants regarding their AL that could help arrive at conclusions about their knowledge of language assessment as LAL is closely and inevitably related to AL, and cannot be considered completely separate from it.

3.4.2 Administration of the Measure

The participants completed the measure on different sessions that were arranged at the end of some of their program courses. The sessions were proctored and the participants were given approximately 45 minutes to complete the measure. Before the start of the sessions, the participants were reminded that the participation was on a voluntary basis, and they were provided with a debriefing form (Appendix B) and an informed consent form (Appendix C).

The modified CAK consisted of 27 multiple-choice questions, and each item had four options, one of them being the correct answer and three being the distractors. The participants were also required to answer 6 questions on the same form that aimed to collect information regarding their backgrounds. The background questions asked participants about their age, gender, CGPA, prior attendance to a workshop or seminar on assessment, their perception of level of preparedness for the overall job of being a classroom teacher, and their perception of level of preparedness for assessing student performance. LAL knowledge base part of the measure was arranged in three different versions with shuffled orders of the items in each version in order to avoid any bias that could be caused by (a) an ordering effect, (b) tiredness effect or (c) exam cheating. The reordering of each version was made through placing each scenario in different orders.

3.4.3 Development and Adaptation of the Measure

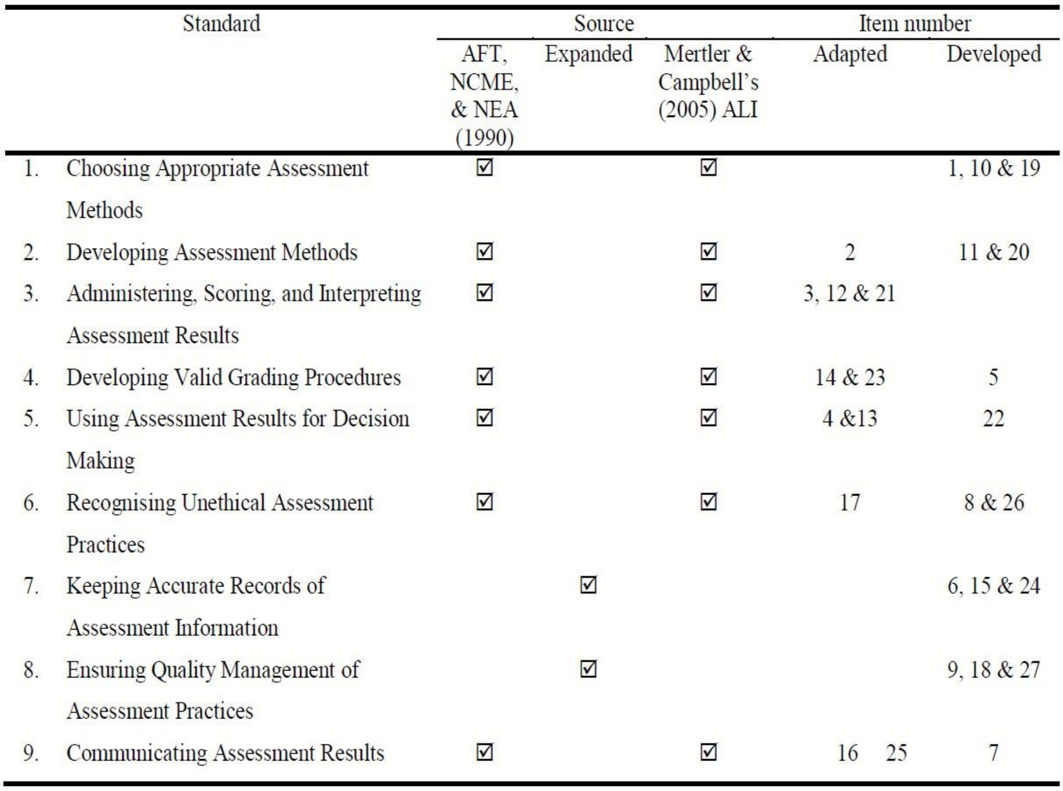
A large number of measures have so far been developed and designed in order to investigate the knowledge base of teacher AL, including self-reported or self-evaluation scales and more objective ones. Because the main focus of the present study was to explore the psychometric properties of a measure to assess pre-service EFL teachers' AL at the knowledge base, the study put more emphasis on more objective measures. There are some very extensively-used measures with reports of high psychometric qualities. These measures, although they contain base or sample questions, are usually adapted to specific contexts before they are used in research studies (Xu & Brown, 2016). Among such measures are Teacher Assessment Literacy Questionnaire (TALQ; Plake, 1993), Classroom Assessment Knowledge Inventory (CALI; Mertler, 2003), and Assessment Literacy Inventory (ALI; Mertler & Campbell, 2005). In fact, CALI is a revised version of TALQ, and both of the measures are aligned with the standards set by the *Standards* document

(AFT, NCME, NEA, 1990). ALI is also based on the *Standards* document, but it is thought to have a more user-friendly format as it contains a contextualised series of items related to a single scenario rather than out-of-context separate individual items.

For the present study, the researcher decided to use a modified version of a measure called CAK adapted by Tao (2014), mainly from ALI, as it has a more user-friendly format and the author reported statistically significant psychometric qualities (X^2=0.68, DF= 2, GFI=1, AGFI=1, RMSEA=0.01, CFI=1).

The original scale was designed to test the AL knowledge base of EFL instructors working at an English-major department and an English non-major department at a higher education setting. It consisted of a total of 27 items, all of which were in the multiple-choice question format, each with four options. The items were designed to correspond to nine standards of teacher competencies for the educational assessment of student performance. Each standard was represented by three items. Appendix (C) presents the original measure. Seven of the standards were taken from the Standards document. The author, having conducted an "extensive review of the existing literature" (Tao, 2014, p. 103) added two more standards, taking into consideration criticisms associated with the narrow aspects of the original standards" (p. 107). The criticisms were especially focussed on activities related to classroom assessment that teachers are required to do in their day-to-day instruction. The added two standards (Standards 7 and 8, as shown in Figure 1) were related to keeping accurate records and managing quality assurance.

The original measure adapted eleven of the items (Items 2, 3, 4, 12, 13, 14, 16, 17, 21, 23, and 25, see Appendix D) from ALI (Mertler & Campbell, 2005), while the rest of the items were developed from scratch following a comprehensive review of literature. Figure 1 shows the standards and the corresponding items in the original measure.



Standard	Source			Item number	
	AFT, NCME, & NEA (1990)	Expanded	Mertler & Campbell's (2005) ALI	Adapted	Developed
1. Choosing Appropriate Assessment Methods	☑		☑		1, 10 & 19
2. Developing Assessment Methods	☑		☑	2	11 & 20
3. Administering, Scoring, and Interpreting Assessment Results	☑		☑	3, 12 & 21	
4. Developing Valid Grading Procedures	☑		☑	14 & 23	5
5. Using Assessment Results for Decision Making	☑		☑	4 &13	22
6. Recognising Unethical Assessment Practices	☑		☑	17	8 & 26
7. Keeping Accurate Records of Assessment Information		☑			6, 15 & 24
8. Ensuring Quality Management of Assessment Practices		☑			9, 18 & 27
9. Communicating Assessment Results	☑		☑	16 25	7

Figure 1: The Standards and Corresponding Items on the Original Measure (Taken from Tao, 2014, p. 108)

The original measure, before being used in the current study, went through an extensive adaptation procedure, including the reordering of some items in order to follow the order of the standards (Table 6). The modified version was firstly adapted to the Turkish context. The adaptation process involved obtaining feedback, revisions and suggestions from a panel consisting of 12 specialists in English Language Teaching, English Language Testing and psychometrics. The panellists were sent the measure and asked to (a) specify whether they think each item on the measure is able to address and assess the related sub-component (standard) of AL, (b) check the answer key and make sure that each item has only one correct option, and (c) provide any comments and/or suggestions to improve each item based on considerations relating but not limited to content, accuracy and wording, using a form to evaluate the measure (Appendix E) they were provided with.

Table 6: Reordering of the Items on the Measure

Original CAK	Modified CAK
1	Q1
2	Q2
3	Q3
4	Q5
5	Q22
6	Q7
7	Q9
8	Q6
9	Q8
10	Q10
11	Q11
12	Q12
13	Q14
14	Q13
15	Q16
16	Q18
17	Q15
18	Q17
19	Q19
20	Q20
21	Q21
22	Q23
23	Q4
24	Q25
25	Q27
26	Q24
27	Not used

Based on the feedback provided by the panellists, a large number of revisions were made on both the content and wording of the items**.** The panel suggested removing Item 27 on the grounds that all of the options could possibly be considered to be correct. A new item (Q26 in the modified version) was created and used in the study. Having made the due revisions and changes, the researcher came up with an adapted and modified version of the original measure (see Appendix F). The standards and items associated with them are presented in Table 7. The 27 items on the measure corresponded to the nine standards (each standard represented by three items) based on three scenarios in total. In other words, the three scenarios each used by the measure contains nine items, each addressing one of the standards for teacher competencies for the educational assessment of student performance.

Table 7: Items on the Modified Measure and Corresponding Standards

Standard	Item
1. Choosing Appropriate Assessment Methods	1-10-19
2. Developing Assessment Methods	2-11-20
3. Administering, Scoring, and Interpreting Assessment Results	3-12-21
4. Developing Valid Grading Procedures	4-13-22
5. Using Assessment Results for Decision Making	5-14-23
6. Recognising Unethical Assessment Practices	6-15-24
7. Keeping Accurate Records of Assessment Information	7-16-25
8. Ensuring Quality Management of Assessment Practices	8-17-26
9. Communicating Assessment Results	9-18-27

3.5 Data Analysis

The current study’s Research Question 1 and Research Question 2 aimed to investigate (a) the psychometric properties of an instrument that could be used to assess EFL teachers’ AL knowledge base, and (b) the current AL knowledge base of pre-service EFL teachers in the Turkish context. The researcher attempted to find answers to these two questions using various analytical techniques including 1PL IRT model, Rasch analysis, CTT methodology, 2PL IRT model, and Rasch PCA. In addition, in order to address Research Question 3, which inquired what factors, if any, influenced LAL, inferential statistics (i.e., Pearson and Spearman Correlations) were used. Table 8 presents a summary of the analytical techniques used by the study in order to answer the research questions.

Table 8: Analytical Techniques Used in Data Analysis

	Method	Research Question
1	1PL IRT model	RQ 1 & RQ 2
2	Rasch Analysis	RQ 1 & RQ 2
3	CTT Methodology	RQ 1 & RQ 2
4	2PL IRT Model	RQ 1 & RQ 2
5	Rasch PCA	RQ 1
6	Descriptive Statistics	RQ 2
7	Correlation	RQ 3

3.5.1 Rasch Analysis

Rasch Analysis was used in the study, as it provides a comprehensive and extensive set of information on the psychometric characteristics of tests and test items, in line with the aim of Research Question 1 and Research Question 2. Rasch Analysis shares a number commonalities with IRT family statistics, and was designed to tackle some problems associated with CTT methodology (Bond & Fox, 2015). Rasch model is able to demonstrate information regarding the difficulty of items together with potential factor structure of the measure. It also provides a comprehensive overview of the test and test items to make sense of the data in the first place (Wright & Stone, 1979).

The creation of interval scale for not only item difficulty but also person ability is allowed in Rasch model, which makes it possible to have a look at both how the items work relative to one another, and how the persons perform relative to other test-takers and the difficulty levels of the items. Logits are used in the reporting of Rasch scores, and they are placed on a scale that measures both item difficulty and person ability (Andrich, 2004). In Rasch model, the probability that a person will correctly answer an item and the probability that an item will be correctly answered by a person at a certain ability level are calculated. If the observed data produces results that are not expected by the model, it may mean that there is a misfit between the model and the data (Wright & Stone, 1979).

Indices like item reliability, separation, fit and thresholds in Rasch Analysis are used to inspect an assessment instrument's psychometric properties such as validity and reliability. An item's ability to be replicated based on the estimates of the Rasch model is represented by item reliability (Bond & Fox, 2015). The closer a value to 1, the more reliable an item is considered, whereas the closer a value to 0, the less reliable the item may be in the sense that there are problems with regard to certainty of replicating the item based on the estimates of difficulty. In terms of separation, which is related to item difficulty variations, greater values refer to better distribution of item difficulties (de Ayala, 2013).

It is also possible in Rasch Analysis to investigate item and person fit statistics (infit and outfit) to detect responses that are problematic. Discrepancies between item responses are represented by infit and outfit (de Ayala, 2013), where infit is weighted by values that are close to the expected difficulty or ability value, with outfit being unweighted, making it more sensitive to responses that are outlying. If persons at a high ability level cannot get easy items, this could lead to infit violations, and outfit violations can appear when an item difficulty is placed outside the response patterns (Linacre, 2000). Mean Square values (MNSQ) and standardised *z*-values (ZSTD) are used to report infit values, which respectively indicate the distortion amount and model-fit unlikelihood.

3.5.2 CTT Methodology

Although Rasch Analysis, 1PL and 2PL IRT models were used for the psychometric investigation of the modified CAK and its items, the researcher decided to employ the traditional CTT methodology as well in order to cross-check the data results produced by the two distinct approaches to the Measurement Theory.

The two approaches to the Measurement Theory have been seen as rivals for a very long time, and the use of IRT-based models has grown exponentially for the past decades, often motivated by the assumption that CTT approach is relatively weaker compared to the IRT approach (Fan, 1998). However, such an assumption could be misleading, considering that both approaches have their advantages and disadvantages. Fan (1998), who presented an important empirical comparison of the two approaches, also provided a summary of the relationship between the advantages and disadvantages of the two approaches. CTT, being too much dependant on the group of test-takers, and relying on information provided from the group as to the entirety of the test (which is considered to be a weakness) has a major advantage of possessing relatively weak assumptions. IRT, on the other hand, does not depend on the group-test relationship, but it comes with the cost of having very strong assumptions. Departing from this complex relation between the two approaches, Fan (1998) compared them empirically through a test battery consisting of 60 math and 48 reading items taken by 193,000 test-takers. The results showed that CTT person statistics were "highly comparable with those" of the three IRT models he used (p. 14). Other results that were comparable in both approaches were related to item difficulty indices. Item discrimination indices, although not as highly comparable as person statistics or item difficulty indices, were also moderately high to highly comparable in both approaches. Therefore, it was considered that performing traditional item and test analyses could provide some insight into the psychometric investigation of the measure.

The origins of the CTT go back to the work of Spearman early in the 20th century (Szabó, 2012). The starting point of an age-old battery of methods in CTT was his *The True Score Model*, which is based on the following formula:

$X=T+E,$

where X is a person's observed score, which is the sum total of the true score (T) and the measurement error (E).

CTT may provide insight into test reliability as well as item and test characteristics including item difficulty and discrimination index. Item difficulty value is simply acquired by calculating the percentage of persons getting an item correctly. And one common way of determining item discrimination in CTT is the subtraction of the number of correct answers in the bottom group from the number of correct answers in the top group, and dividing it by the number of persons in the top group. The literature usually advises the following guidelines as to the interpretation of the discrimination index (*D*) in CTT (Szabo, 2012, p. 32):

1. If $D \geq .40$, the item is functioning quite satisfactorily
2. If $.30 \leq D \leq .39$, little or no revision is required
3. If $20 \leq D \leq .29$, the item is marginal and needs revision
4. If $D \leq .19$, the item should be eliminated or completely revised.

3.5.3 IRT

IRT, also referred to as latent trait theory, was developed in the 1960s in order to expand on (not to replace) CTT by overcoming some problems associated with the latter (Szabó, 2012). In IRT models, it is considered that the relationship between the difficulty of an item and the ability of a person is the primary factor in determining how likely a certain person is to get an item correctly, thus making the concepts of ability and item difficulty central to IRT models. This relationship is described with an ICC (Figure 2).

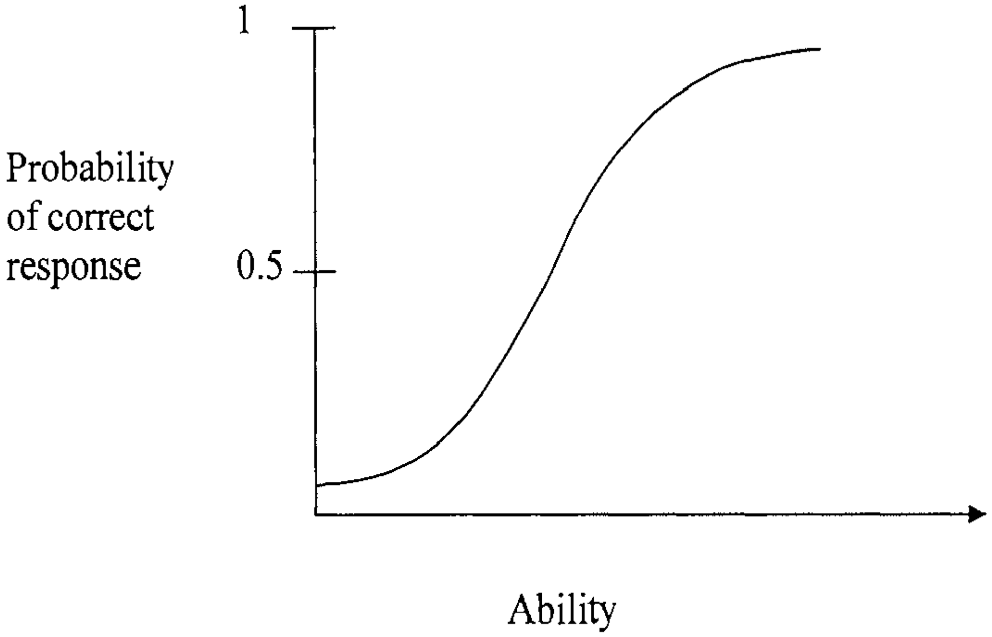


Figure 2: An Example ICC (Szabó, 2012)

The relationship between ability and difficulty can be usually visually reviewed with an ICC, drawing conclusions about the difficulty level of an item. The ICC shape provides information also on the discriminatory power of an item: the steeper it is the greater the discrimination.

The two main assumptions of IRT models are model-fit and unidimensionality. The former relates to a requirement that there is a good fit between the particular model and the data, and vice versa, while the latter requires that a set of items on an assessment instrument assesses only one latent trait.

Although there are several IRT-based models for item analysis, the three groups of models 1PL, 2PL, and 3PL are among the most widely-used models in educational measurement practices. These models are distinguished by the number of parameters they address and their statistical assumptions. All three models have at least one parameter besides the person ability parameter; that is, 1PL model has one parameter, item difficulty, in addition to person ability. 2PL has the item discrimination parameter in addition to item difficulty and person ability. Similarly, besides the parameters of person ability, item difficulty and item discrimination, 3PL model has the pseudo-chance level parameter, which refers to the likelihood of a person to get an item right with sheer guessing. These parameters are represented by the letters a, b, and c, as shown below:

parameter *a*: discrimination
parameter *b*: item difficulty
parameter *c*: pseudo-chance level

Choosing the appropriate IRT model is one of the most-widely encountered problems for researchers. The decision is often made according to considerations regarding the model fit, complexity of the analysis and sample size. Because 3PL model usually fits with large sample sizes, 1PL and 2PL were used for the examination of the psychometric characteristics of the measure used in the present study.

3.5.4 Rasch PCA

Rasch PCA was performed as part of the analyses of the psychometric properties of the modified CAK. Similar to other Rasch analytical procedures, Rasch PCA looks at the differences between modelled prediction and the observed data in order to look at the data patterns using residuals. More precisely, the purpose of Rasch PCA was to test dimensionality in the data. Items having similar data patterns share a substantive attribute, and thus may create a component or a dimension. The technique helps identify properties shared by items on the measure. It aims to discover the measure’s structure with the help of standardised residuals (Linacre, 1998). Using this approach, it may hint at the presence of secondary components or sub-dimensions in an assessment tool. Unlike Confirmatory Factor Analysis, Rasch PCA is not used with the aim of testing theories or hypotheses, but of exploring or describing the relationships among item groups.

Reduction of the dimensions present within the data is the primary goal of conducting a Rasch PCA. Linacre (1998) states that the amount of variance present in the variables of an assessment instrument is reduced in Rasch PCA into a smaller group of variables. In other words, the amount of possible correlations among variables is transformed into a smaller set of uncorrelated variables referred to as principal components (Wright, 1996). The first principal component uncovered explains the largest amount of variance within the data, and each subsequent component goes on to account for as much variance as possible. Each of the components extracted represents a separate component or construct. Rasch PCA assigns a factor loading value to each item. Factor loadings represent the correlation coefficients between the factors and variables within the data. The groupings of the items based on the factor loadings are used to understand the underlying nature of certain dimensions of the instrument (Bond & Fox, 2015). Furthermore, Rasch PCA uses eigenvalues in order to describe the amount of variance explained by the data. Any existing component should have an eigenvalue above 2 to suggest a presence of a strong dimension within the variables (Bond & Fox, 2015).

CHAPTER 4

FINDINGS AND DISCUSSION

Investigation of the modified Classroom Assessment Knowledge (CAK) measure was the first research objective of the present study in line with the Research Question (1): What are the psychometric properties of a language assessment literacy knowledge scale adapted from the Standards for Teacher Competence in Educational Assessment of Students (AFT, NCME, & NEA, 1990) and Mertler (2003)? The psychometric properties inquired by the study included item and test difficulty levels, item discrimination, content and construct validity, and internal consistency reliability as well as item and test information functions. In order to obtain a more complete picture regarding the psychometric properties, 1PL model, Rasch model, CTT methodology and 2PL model were employed, as each of these approaches deals with more or less the same questions through different perspectives.

4.1 Findings

This section presents the findings from the analytical techniques used to explore the psychometric properties of the modified CAK measure. The results from IRT models, Rasch model and CTT are given in the following paragraphs.

4.1.1 1PL Model

The first round of the analyses started with 1PL analysis, which is among the latent variable models constituting a general class of models used to analyse multivariate data, which was performed using the *ltm* package in R, a free programming language for statistical computing. The *ltm* package offers item analysis procedures for multivariate dichotomous and polytomous data including Rasch, 2PL, Birnbaum's 3PL, and Samejima's Graded Response model (Rizopoulos, 2006). The purpose of the 1PL analysis was to gain a brief overview of the instrument and its items.

Model fit is an important concept in both IRT models and Rasch model (Bonf & Fox, 2015). It explores whether the data provided for the analysis is suitable for the model. There are a number of criteria for the goodness of fit. For the 1PL analysis Chi-square (x^2) was checked. x^2 is a statistical test of significance used to assess the null hypothesis (i.e., there is no difference between the data and the theoretical model used for the assessment). Depending on the sample size, it explores the overall fit of the model to the data.

In the 1PL analysis, responses from 74 participants to the 27 items on the instrument were examined. In the *ltm* package, the fit of the model to the data is checked with a function called *GoF.rasch*, which tests the null hypothesis through generating *B* samples using likelihood estimates, and the Person's x^2 statistics. T_b is calculated for each data set, after which the *p*-value is approximated by the number of times $T_b \geq T_{obs}$ plus one, which is divided by *B*+1, where T_{obs} corresponds to the value of the statistic in the original data set (Rizopopulos, 2006).

The results from the analysis showed that the model fit was good for the data (T_{obs}: 96203891, #datasets: 50, *p*: .32 [*p*> .05]). Difficulties of the items were investigated, as in the model there is only one parameter (item difficulty) analysed aside from the person ability parameter. In other words, item discrimination is not specifically addressed in the model, where each item is designated a fixed value for item discrimination. As for the item difficulty parameter in the model, a difficulty value of 0 means that the item measures persons with an average ability. And a difficulty value above 0 means that the item measures persons with higher ability, whereas a value below 0 means that the item measures persons with lower ability. Item difficulty levels obtained from the analysis for 27 items on the instrument are shown in Table 9.

Table 9: An Overview of the Item difficulty Values from 1PL Analysis, with the Fixed Item Discrimination Value Set Equally for Each Item at .68

Item	Difficulty (*b*)	P (x=1 | z=0)
Q10	-3.15	0.90
Q5	-2.33	0.83
Q2	-2.06	0.80
Q19	-1.93	0.79

Table 9 (continued)

Q24	-1.82	0.78
Q9	-1.70	0.76
Q12	-1.48	0.73
Q1	-1.27	0.70
Q15	-0.79	0.63
Q13	-0.70	0.62
Q22	-0.34	0.56
Q7	-0.34	0.56
Q25	-0.25	0.54
Q11	-0.07	0.51
Q6	-0.07	0.51
Q17	-0.07	0.51
Q8	0.10	0.48
Q20	0.36	0.44
Q3	0.45	0.42
Q4	0.63	0.39
Q14	1.00	0.34
Q26	1.19	0.31
Q23	1.39	0.28
Q18	1.50	0.26
Q27	1.82	0.22
Q21	2.07	0.20
Q16	2.47	0.16
Mean	- 0.20	0.53

The items in the table are ordered according to difficulty from the easiest to the most difficult ones. The column P (x=1 | z=0) in the table refers to $P(x_i = 1 \mid z=0)$ under the model fit, and refers to an average person's (a person at an average ability level) probability of getting the *i*th item right. For instance, an average ability person's probability of getting Q16 right is 16% whereas the same person is 90% likely to get Q10 right. As can be seen in the table, the difficulty value of an item negatively correlates with the probability that an average person will get that item right. 16 of the items (Questions 10, 5, 2, 19, 24, 9, 12, 1, 15, 13, 22, 7, 25, 11, 6, and 17) were the easier ones with a difficulty value below zero, whereas 11 of them

(8, 20, 3, 4, 13, 26, 23, 18, 27, 21, and 16) were the more difficult ones with difficulty values above zero. Items with difficulty values close to zero could be said to target persons closer to average ability. With a mean value of - 0.20, the overall measure could be said to have an average difficulty level. In other words, a person at an average ability level is likely to answer approximately half of all questions in the measure correctly. So, this analysis suggests that the difficulty level of the measure appears to be moderate.

Of all the 27 items, Q10 (choosing an appropriate assessment method) was by far the easiest item on the instrument, with a difficulty value of - 3.15. The item corresponded to *Standard 1*, which was related to choosing the appropriate assessment method. On the other hand, the item with the highest difficulty value was Q16, which was related to keeping accurate records of assessment information (*Standard 7*).

ICCs were also obtained for each item in order to provide a visual review of the difficulty levels of the items on the instrument. ICC is a curve providing considerable amount of information on an item in IRT-based models (Baker, 1985). The vertical line represents probability for a person at a certain ability level of getting a certain item right, and the horizontal line represents the ability level. Figure 3 presents ICCs for all the items on the instrument. Because the model does not take item discrimination into account, which would be represented by the steepness of a curve (the steeper the curve, the better the discrimination; and the flatter the curve, the worse the discrimination), the steepness (the discrimination) of the curves for all items are modelled equal, but they differ in their difficulty. The easier items are placed above the 0.5 probability level where the more difficult items are placed below the 0.5 probability level. In the horizontal line, 0 refers to the average ability person, 2 refers to 2 standardised units above average ability, and - 2 stands for 2 standardised units below average ability. IIC was also plotted to display the range of abilities the items on the instrument assess (Figure 4). For instance, Q10 tends to measure persons with ability around 3.5 standardised units below average ability, whereas Q16 tends to measure test-takers with ability around 2.47 standardised units above average ability. And a review of the range of abilities the whole measure targets through the Test Information Function (Figure 5), it could be argued that the curve represents a normal distribution and that the measure contains items that tend to assess persons from all ranges of ability. The package also provides the total information that displays the percentage the measure is able to provide information about across the persons' latent ability. The values for total information are as follows:

Total Information = 18.4,
Information in (-4, 4) = 15.29 (83,07%),

meaning that the instrument with 27 items can effectively provide information for 83,07% of the persons' latent ability. In other words, because test information function has a relationship with standards error, and thus with reliability, these values seem to be denoting a high level of reliability.

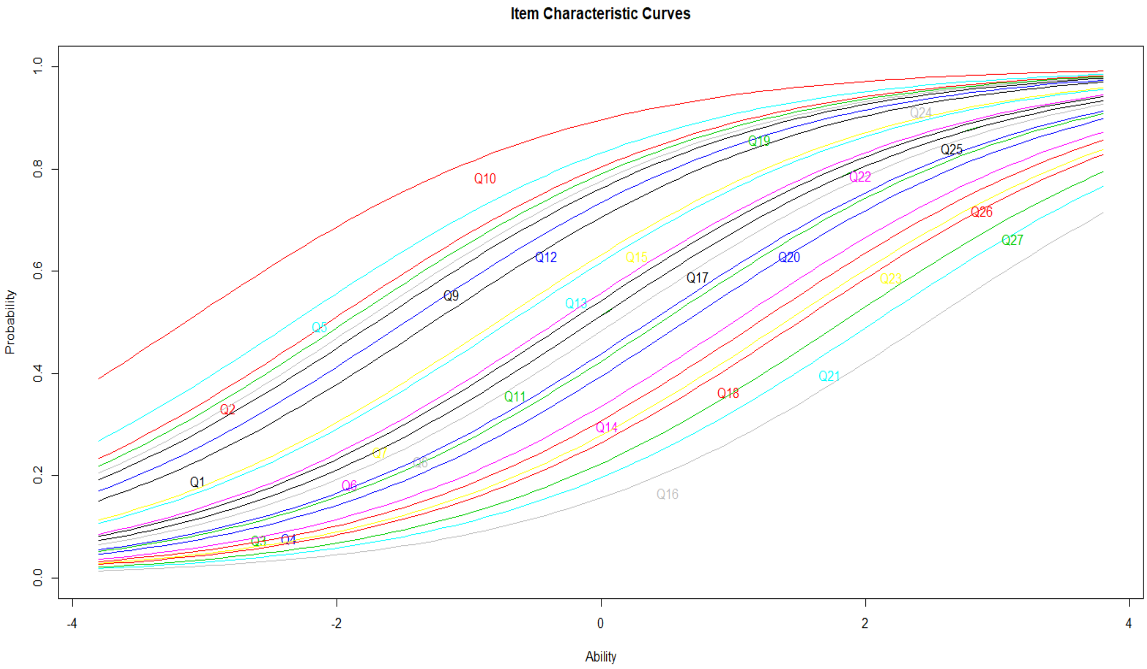


Figure 3: ICCs for All Items

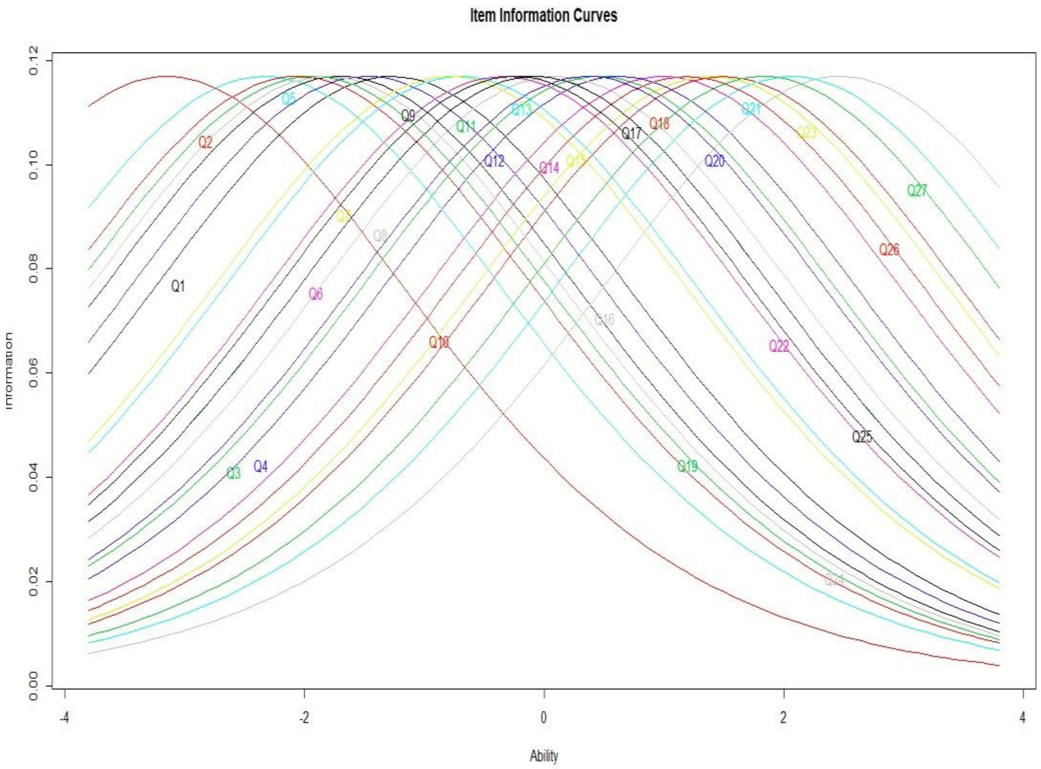


Figure 4: Item Information Curves for All Items

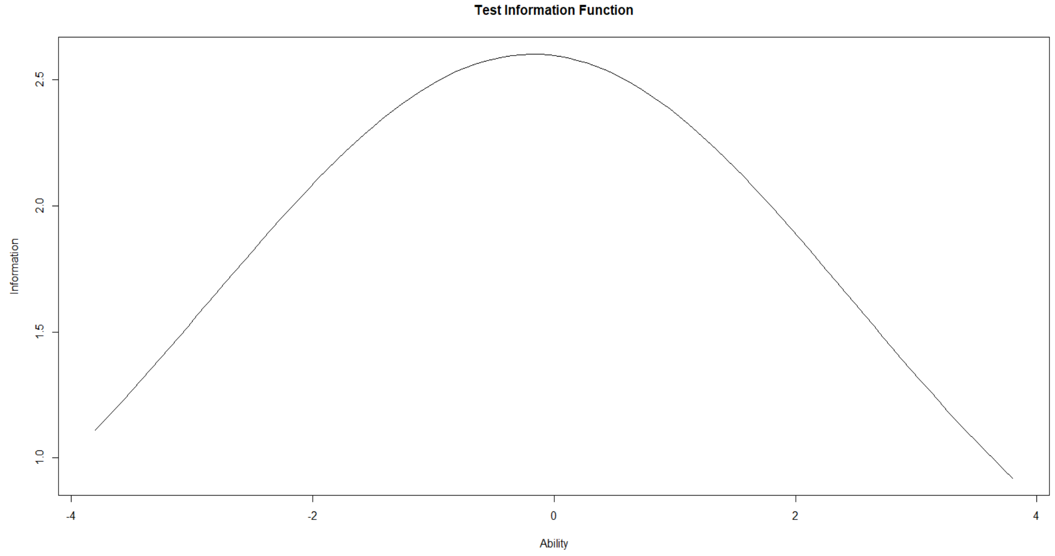


Figure 5: Test Information Function (1PL)

Visual inspection of the figures (Figures 3, 4, and 5) relating to ICCs, IIFs and test information function could provide a substantial amount of insight into the functioning of the items. The ICCs show the different ability levels each item targets, and that there are no or only slight overlaps among items in terms of their targeted ability levels. IICs, on the other hand, seem to support the interpretation of item characteristic curves, providing evidence for the idea that the items on the measure afford information regarding persons at different ability levels, with clustering of matching items and persons. And the test information function represents a pretty much normal distribution, while leaning slightly towards to the left side, meaning that although the measure largely targets persons at average ability, the number of items targeting persons below average ability is slightly more than the number of items targeting persons above average ability, but the difference is not a big one. In other words, it is possible to infer that the overall difficulty of the measure is moderate.

4.1.2 CTT Methodology

1PL model only takes into account the item difficulty parameter in the item analysis, and as fit statistics are not suggested to be the only indicators for deciding whether to accept or reject items as they are dependent on sample size (Bond & Fox, 2007; Wu & Adams, 2007), reliability and discrimination indices from the traditional item analysis can be used as a follow-up inspection in combination with fit statistics to evaluate the psychometric properties of the items and test. Therefore, CTT methodology was employed for item and test analysis using the Test Analysis Program (TAP, version: 19.1.4) made freely accessible by Ohio University at https://people.ohio.edu/brooksg/#TAP. Responses given by 74 participants to the 27 items were examined. The descriptive statistics regarding the traditional item and test analysis are shown in Table 10. The minimum score from the instrument was 1 (3,7%) while the maximum score was 20 (77.8%), meaning that the person with the highest score got around 80 percent of all items correctly.

There was no participant who answered all the questions on the instrument accurately. Mean item difficulty value was 0.53, and mean item discrimination index was 0.36. Cronbach’s alpha (denoting test reliability) value was 0.70, indicating reasonable reliability (Taber, 2018). The closer this value is to 1, the higher the reliability of an assessment instrument is.

Table 10: Descriptive Statistics from the Traditional Item Analysis (N=74)

Number of Items Analysed	24
Total Possible Score	24
Minimum Score	1 (3.7%)
Maximum Score	20 (77.8%)
Median Score	13.51 (58.3%)
Mean Score	13.5. (58.3%)
Standard Deviation	4.10
Variance	16.82
Skewness	- 0.70
Kurtosis	0.25
Mean Item Difficulty	0.56
Mean Discrimination Index	0.40
Mean Point Biserial	0.36
Mean Adj. Point Biserial	0.27
KR20 (Alpha)	0.73
SEM (from KR20)	2.29
High Grp Min Score (N=20)	18
Low Grp Max Score (N=23)	12

Further details regarding the item analysis are provided in the following paragraphs. Table 11 presents an overview of the individual analyses of the items.

Table 11: Results from the Traditional Item Analyses (N=74)

Item	Number Correct	Difficulty	Discrimination	# Correct in High Grp	# Correct in Low Grp
Q1	51	0.69	0.61	20	9
Q2	59	0.80	0.43	20	13
Q3	32	0.43	0.53	14	4
Q4	30	0.41	0.48	14	5
Q5	61	0.82	0.34	19	14
Q6	38	0.51	0.25	12	8
Q7	41	0.55	0.50	16	7
Q8	37	0.50	0.50	16	7
Q9	55	0.74	0.70	20	7

Table 11 (continued)

Q10	66	0.89	0.20	19	18
Q11	38	0.51	0.58	15	4
Q12	52	0.70	0.38	18	12
Q13	45	0.61	0.60	19	8
Q14*	25	0.34	0.05	7	7
Q15	46	0.62	0.27	14	10
Q16*	13	0.18	0.08	5	4
Q17	39	0.53	0.26	11	9
Q18	21	0.28	0.33	10	4
Q19	57	0.77	0.30	19	15
Q20	32	0.43	0.35	13	7
Q21	16	0.22	0.19	6	3
Q22	41	0.55	0.37	16	10
Q23	22	0.30	0.51	12	2
Q24	56	0.76	0.33	18	13
Q25	41	0.55	0.35	13	7
Q26	24	0.32	0.46	11	2
Q27*	18	0.24	-0.02	3	4
Mean	-	0.56	0.40	-	-

* Items indicating problems. They are removed from the test and overall test statistics.

The difficulty index shows what percentage of the test-takers got an individual item correct. The value can be between 0 and 1. The closer the value is to 0, the more difficult it is; and the closer the value is to 1, the easier it is. Table 11 shows that Q16 was the most difficult item, and Q10 was the easiest. On the other hand, the discrimination index shows how successful an item is in terms of discriminating high-achieving test-takers from low-achieving ones. Three questions on the measure (Q14, Q16, and Q27) were shown to have extremely low discrimination indices (Table 12 presents the options analysis regarding these items) and they were eliminated (i.e., removed from the measure, and ignored in the further statistical analyses including Rasch analysis, 2PL IRT analysis, and Rasch PCA). In contrast, aside from five items (Qs 6, 10, 15, 17, and 21) showing relatively lower discrimination power, the rest of the items on the instrument had reasonable or high discrimination power, with Q9 having the highest discrimination index value at .74.

Table 12: Options Analysis for the Eliminated Items

Item		Option A	Option B	Option C	Option D
Q14	TOTAL	19 (0.257)	25*(0.338)	19 (0.257)	9 (0.122)
	High	6 (0.300)	7 (0.350)	3 (0.150)	4 (0.200)
	Low	5 (0.217)	7 (0.304)	7 (0.304)	2 (0.087)
	Diff	1#(0.083)	0 (0.046)	-4(-0.154)	2#(0.113)
Q16	TOTAL	17 (0.230)	13*(0.176)	11 (0.149)	31 (0.419)
	High	6 (0.300)	5 (0.250)	2 (0.100)	7 (0.350)
	Low	6 (0.261)	4 (0.174)	3 (0.130)	9 (0.391)
	Diff	0 (0.039)	1 (0.076)	-1(-0.030)	-2(-0.041)
Q27	TOTAL	2 (0.027)	18*(0.243)	29 (0.392)	24 (0.324)
	High	0 (0.000)	3 (0.150)	12 (0.600)	5 (0.250)
	Low	2 (0.087)	4 (0.174)	5 (0.217)	11 (0.478)
	Diff	-2(-0.087)	-1(-0.024)	7#(0.383)	-6(-0.228)

Item discrimination is often likely to negatively function when an item is answered correctly by more test-takers from the low ability group compared to the high ability group, which indicates that the item is not able to differentiate between the high-achieving and low-achieving test-takers. Table 13 shows the content of the eliminated items and the standards they belong to.

Table 13: Contents of the Eliminated Items

Item	Scenario	Content	Standard
Q14	*Ms. Zeynep Demir is an EFL teacher working in a high school setting. She has just finished teaching a unit on climate change and wishes to measure her students' understanding of this particular unit using a multiple-choice test where each item has only one correct option.*	*Some of Ms. Demir's students do not score well on the multiple-choice test. She decides that the next time she teaches this unit, she will begin by administering a pretest to check for students' prerequisite knowledge. She will then adjust her instruction based on the pretest results. What type of information is Ms. Demir using?* *A) Norm-referenced information* *B) Criterion-referenced information* *C) Both norm- and criterion-referenced information* *D) Neither norm- nor criterion-referenced information*	5
Q16	*Ms. Zeynep Demir is an EFL teacher working in a high school setting. She has just finished teaching a unit on climate change and wishes to measure her students' understanding of this particular unit using a multiple-choice test where each item has only one correct option.*	*Ms. Demir understands that her classroom assessment records serve the following purposes except ----.* *A) provide an overview of assessment methods developed* *B) demonstrate diagnostic information regarding the students* *C) show the extent of student progress throughout the instruction* *D) inform administrative decision makers on various issues*	7

Table 13 (continued)

Q27	*Mr. Ahmet Kaplan is a senior EFL lecturer in a higher education setting. Experienced in issues of classroom assessment, Mr. Kaplan is often asked by his colleagues to respond to questions concerning best practices for evaluating student learning.*	*A student in Mr. Kaplan's class receives a raw score of 12 items answered correctly out of a possible score of 15 on the vocabulary section of a test. This raw score equates to a percentile rank of 45. He is confused about how he could answer so many items correctly, but receive such a low percentile rank. He approaches Mr. Kaplan for a possible explanation. Which of the following is the appropriate explanation to offer to the student?* *A) "I don' know… there must be something wrong with the way the test is scored. I'll check immediately."* *B) "Although you answered 12 correctly, numerous students in the class answered more than 12 correctly."* *C) "Raw scores are purely criterion-referenced, but percentile ranks are merely one form of norm-referenced scoring."* *D) "Raw scores are purely norm-referenced, but percentile ranks are merely one form of criterion-referenced scoring."*	9

4.1.3 Rasch Analysis

With the three problematic items removed from the measure, Rasch analysis was performed on the remaining 24 items responded by 74 participants using Ministeps version of Winsteps®, as this software provides practical item-person maps (Wright maps) and a practical interface for fit statistics (i.e., infit and outfit) to investigate possible problematic items

Central to Rasch analysis are concepts of reliability, separation, logits in relation to item difficulty and person ability, and fit statistics including infit and outfit, all of which together provide insights into the psychometric properties of an assessment instrument. In terms of reliability, values closer to 1 suggest stronger reliability, and values closer to 0 represent weaker reliability in relation to replicating item difficulty estimates. Analysis of 24 items on the instrument found 0.91 reliability value for items, and 0.72 for persons, implying a stronger level of reliability for the replication of the measure for the same group of persons. Separation denotes the variation within item difficulties and person abilities. Separation values were 1.61 for persons, and 3.25 for items respectively, meaning that there were not even two groups of person ability but that the items could be divided into at least three groups based on their difficulty (Table 14). Separation value for persons could suggest that the overall sample exhibited a homogenous level of performance in terms of their assessment knowledge.

Table 14: Summary of Rasch Person and Item Statistics

	Total	Count	Infit		Outfit		Reliability	Separation
PERSON								
(Mean)	13.5	24	1.00	.1	.97	.0	0.72	1.61
(SD)	4.1	.0	.15	.8	.26	.8		
ITEM								
(Mean)	41.7	74	1.00	.1	.97	-.1	0.91	3.25
(SD)	13.3	.0	.10	.8	.17	.9		

Rasch analysis provides information regarding point-measure correlations, and they indicate the correlations expected by the model and the correlations that are observed (Linacre, 2012). Positive correlations are desired as they suggest that correct responses to items have positive correlations with the person measures. If the observed correlation is greater than the expected correlation, it shows that the item is over-discriminating between high-achieving persons and low-achieving persons. If the observed correlation is less than the expected correlation, it suggests that the item is under-discriminating between high-achieving and low-achieving persons. The observed and expected point-measure correlations are shown in Table 15. Rasch analysis also provides a review of item and person fit statistics, which helps locate poor fit between the observed data and the model (deAyala, 2013). Infit and outfit statistics, which relate to discrepancies between responses from persons (infit is weighted by the approximity to the expected value of ability or difficulty; and outfit is unweighted, and thus is sensitive to outlier responses), are reported as Mean Square values (MNSQ) and standardised *z*-values (ZSTD). MNSQ values, which demonstrate the amount of distortion present with 1.0 as the expected value, show observations that are predictable if they are above 1.0, and show that observations are unpredictable if they are below 1.0. ZSTD values, which can be either positive or negative, indicate the unlikelihood of the model-data fit with 0.0 as their expected values. To be more specific on MNSQ values, Linacre (2012) suggests the following regarding the interpretation of infit and outfit mean-square fit statistics:

> *>2.0: Distorts or degrades measurement system*
> *1.5-2.0: Unproductive for construction of measurement, but not degrading*
> *0.5-1.5: Productive for measurement*
> *<0.5: Less productive for measurement, but not degrading. May produce misleadingly good reliabilities and separations*

MNSQ values for items on the instrument in addition to their logits are also presented in Table 15. It can be understood from the table that the majority of items have observed correlation values close to the expected correlation values, suggesting that there was no serious problem regarding item discrimination. As for the fit statistics, MNSQ values for all of the remaining items on the measure varied between 0.74 and 1.14 (Q9 and Q19) for infit statistics and between 0.6 and 1.29 (Q9 and Q18) for outfit statistics, all within the acceptable (productive) ranges of 0.5 and 1.5, with a mean infit MNSQ value of 1.0 and a mean outfit MNSQ value of 0.97. The results of the fit statistics suggest that the data obtained from the measure was a good fit for the model.

Table 15: Rasch Analysis Item Statistics

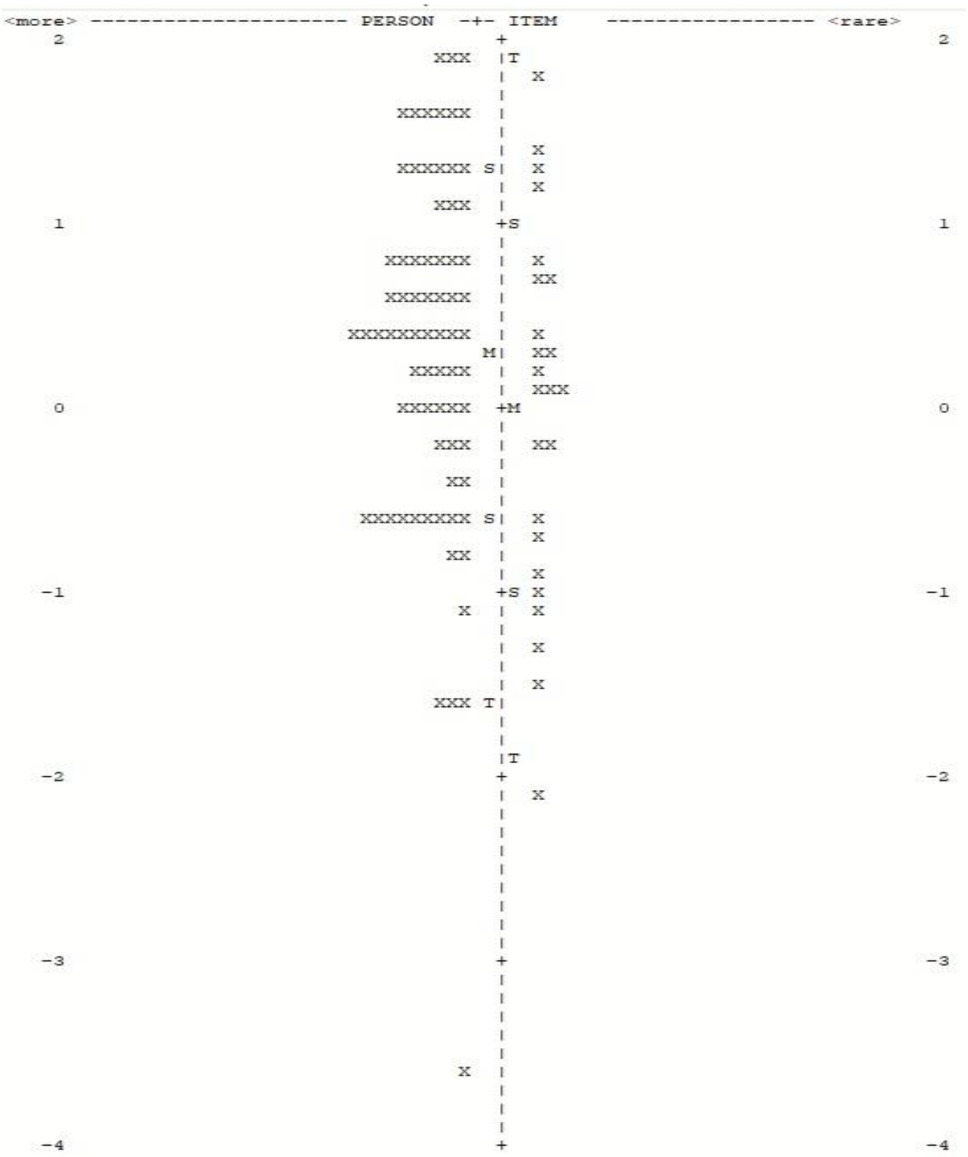
Item	**Logit**	**Infit**		**Outfit**		**PT-MSR**	
		MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.
Q1	-0.58	0.89	-0.89	0.78	-1.25	0.51	0.39
Q2	-1.26	0.9	-0.53	0.76	-0.9	0.48	0.37
Q3	0.67	0.96	-0.49	0.91	-0.55	0.42	0.37
Q4	0.8	0.92	-0.9	0.89	-0.59	0.44	0.36

Table 15 (continued)

Q5	-1.46	0.86	-0.67	0.7	-0.99	0.51	0.36
Q6	0.29	1.04	0.45	1.02	0.16	0.35	0.38
Q7	0.1	1	0.03	0.97	-0.16	0.39	0.38
Q8	0.35	1.02	0.22	1.01	0.13	0.36	0.38
Q9	-0.9	0.74	-1.92	0.6	-2.13	0.64	0.38
Q10	-2.11	1.12	0.48	1.02	0.19	0.26	0.33
Q11	0.29	0.91	-1	0.9	-0.7	0.46	0.38
Q12	-0.66	1.09	0.75	1.1	0.56	0.3	0.38
Q13	-0.16	0.98	-0.17	0.96	-0.26	0.41	0.39
Q15	-0.23	1.12	1.19	1.2	1.28	0.26	0.39
Q17	0.23	1.11	1.29	1.1	0.76	0.28	0.38
Q18	1.42	1.08	0.66	1.29	1.17	0.21	0.32
Q19	-1.07	1.14	0.87	1.14	0.64	0.25	0.38
Q20	0.67	1.07	0.85	1.03	0.26	0.31	0.37
Q21	1.83	1.14	0.9	1.22	0.75	0.15	0.29
Q22	0.1	1.11	1.18	1.1	0.73	0.28	0.38
Q23	1.35	0.92	-0.68	0.83	-0.7	0.42	0.33
Q24	-0.98	1.02	0.16	1.03	0.19	0.36	0.38
Q25	0.1	1.01	0.1	1	0.07	0.38	0.38
Q26	1.21	0.95	-0.48	0.85	-0.68	0.41	0.34
Mean	.01	1.0	1.0	.97	-.1	-	-
SD	.97	.10	.8	.17	.8	-	-

The scores from Rasch analysis are reported in logits that are located on a scale showing both item difficulty and person ability. Logits denote the probability of a person at a certain ability level getting an item at a certain difficulty level right. The results show that the most difficult item was Q21 while Q10 was the easiest item on the measure. It is also possible to have a visual inspection of the fit of the item difficulty and person ability through an item-person map (Wright Map, in Figure 6). Figure 6 shows the positioning of items and persons on a continuum of the latent variable. On the left side are the logits, and the items are placed on the right side. Items become more difficult upwards and easier downwards on the scale. Similarly persons are placed on the left side according to their ability. The range of persons was between -3.55 and 1.89, while the range of items was between -2.11

and 1.83. The vast majority of items and persons were positioned between -1.50 and 1.50 logits, but there were some persons and items that fell outside this range. There were three persons above the logit of the item with the highest logit (Q21), meaning that a slightly more difficult item was needed to match the ability of these persons. Similarly, there was one person at -3.55 logit much beyond the easiest item (Q10) located at -2.11 logit.



```
<more> --------------------- PERSON -+- ITEM    ----------------- <rare>
    2                                   +                                   2
                                   XXX  |T
                                        |  X
                                        |
                                XXXXXX  |
                                        |
                                        |  X
                                XXXXXX S|  X
                                        |  X
                                   XXX  |
    1                                   +S                                  1
                                        |
                               XXXXXXX  |  X
                                        |  XX
                               XXXXXXX  |
                                        |
                            XXXXXXXXXX  |  X
                                       M|  XX
                                 XXXXX  |  X
                                        |  XXX
    0                           XXXXXX  +M                                  0
                                        |
                                   XXX  |  XX
                                        |
                                    XX  |
                                        |
                             XXXXXXXXX S|  X
                                        |  X
                                    XX  |
                                        |  X
   -1                                   +S X                               -1
                                     X  |  X
                                        |
                                        |  X
                                        |
                                        |  X
                                   XXX T|
                                        |
                                        |
                                        |T
   -2                                   +                                  -2
                                        |  X
                                        |
                                        |
                                        |
                                        |
                                        |
                                        |
                                        |
                                        |
   -3                                   +                                  -3
                                        |
                                        |
                                        |
                                        |
                                        |
                                     X  |
                                        |
                                        |
                                        |
   -4                                   +                                  -4
```

Figure 6: Item-person Map (N=74)

4.1.4 2PL Model

In order to provide a latent model perspective into item psychometric properties aside from 1PL and Rasch models, with particular focus on item discrimination, a 2PL item analysis was performed as the data had a good fit with the model (Table 16) following a piecewise goodness of fit check (Reise & Revicki, 2015).

Table 16: 2PL Goodness of Fit Analysis

Item	X^2	*p*
Q1	10.49	0.15
Q2	4.29	0.79
Q3	9.86	0.43
Q4	13.32	0.11
Q5	8.04	0.29
Q6	15.34	0.12
Q7	9.67	0.29
Q8	17.18	0.02
Q9	4.26	0.63
Q10	9.69	0.26
Q11	2.88	0.96
Q12	4.55	0.83
Q13	8.41	0.49
Q15	5.95	0.68
Q17	7.20	0.60
Q18	10.30	0.25
Q19	6.74	0.59
Q20	6.06	0.79
Q21	8.30	0.46
Q22	2.92	0.93
Q23	8.89	0.37
Q24	6.15	0.58
Q25	10.63	0.26
Q26	10.94	0.19

The values for total information in the 2PL model are provided below:

Total Information = 21.27,
Information in (-4, 4) = 19.45 (91.46%),

meaning that the measure with 24 items can effectively provide information for 91.46% of the persons' latent ability. Following the goodness of fit test, the items were analysed for their psychometric properties (i.e., item difficulty and item discrimination). In the two-parameter logistic model, the two parameters are parameter *b* (difficulty) and parameter *a* (discrimination). Baker (1985) provides some insightful information on how to interpret these two parameters. The following description is given for the interpretation of parameter *a* (Baker, 1985, p.34):

Verbal label	**Range of values**
none	0
very low	.01 - .34
low	.35 - .64
moderate	.65 - 1.34
high	1.35 - 1.69
very high	>1.70
perfect	+ infinity

However, there is no such description provided for parameter *b* as it would pose some theoretical problems. Such descriptions as *difficult* or *easy* in CTT methodology denote some comparisons between the groups of test-takers relative to each other. Because IRT models are not group dependent, an item's difficulty is defined as a point on the ability scale, where the probability of a correct response to an item is placed at 0.5 probability for 1PL and 2PL models. Therefore, a proper way of interpreting the difficulty of an item under IRT models could be with respect to where the item functions on the ability scale. For example, an item with a difficulty value of -1.0 functions among lower-achieving persons while an item with a difficulty value of 1.0 functions among higher-achieving persons. Table 17 presents the values for *b* and *a* parameters of the items (N=24) on the measure.

Table 17: Results from the 2PL Analysis

Item	Difficulty (*b*)	Discrimination (*a*)
Q1	-0.74	1.47
Q2	-1.15	1.58
Q3	0.35	0.98
Q4	0.48	0.98
Q5	-1.21	1.81
Q6	-0.07	0.80
Q7	-0.32	0.74
Q8	0.08	0.94
Q9	-0.73	3.03
Q10	-2.65	0.84
Q11	-0.05	1.02
Q12	-2.41	0.40
Q13	-0.66	0.74
Q15	-1.48	0.34
Q17	-0.10	0.52
Q18	3.23	0.29
Q19	-5.11	0.24
Q20	0.52	0.44
Q21	7.06	0.18
Q22	-0.53	0.43
Q23	0.82	1.49
Q24	-1.65	0.77
Q25	-0.23	0.78
Q26	1.14	0.72
Mean	-0.21	0.87

A scatter plot chart based on the results in Table 17 was created in order to have a visual aid to better understand the distribution of item difficulty and discrimination values, and the interaction between them (see Figure 7). As can be seen in the figure, the majority of the items are distributed between 0 amd 1.5 ability level and -2 and 2 difficulty level, denoting normal distribution.

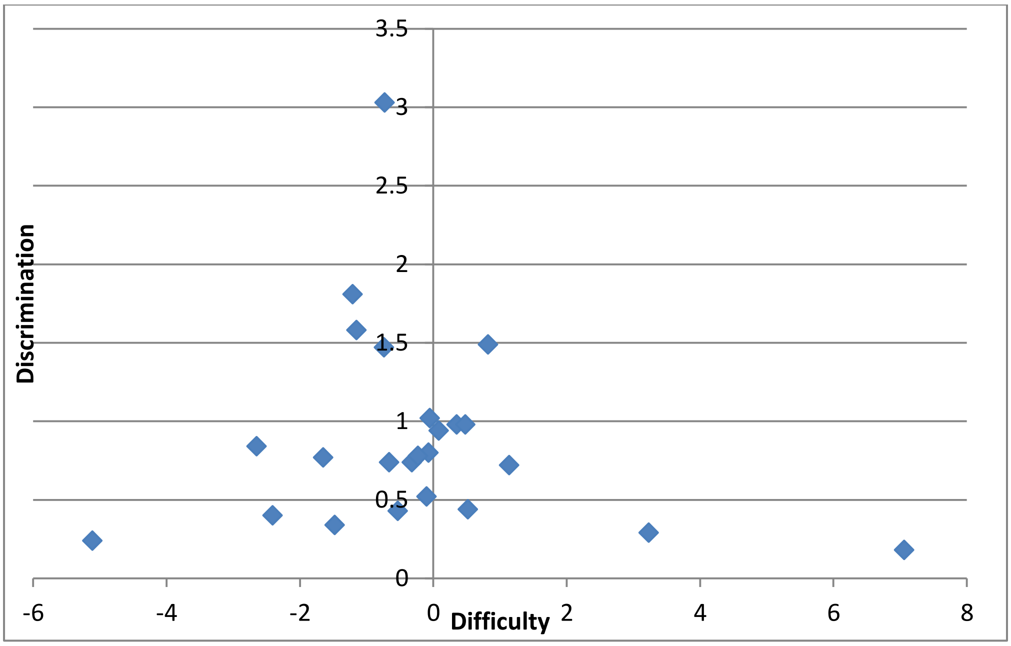


Figure 7: Scatter Plot for Item Difficulty and Discrimination

According to the results of the 2PL analysis, three items (Qs 21, 18 and 19) had very low discriminating power, whereas the rest of the items could be said to have acceptable discrimination indices. The mean value for parameter *a* was 0.87, falling within the range of moderate item discrimination. Q21 was the most difficult item on the measure. A visual analysis of the ICCs and Test Information Function (Figures 8 and 9) also suggests that the majority of the items had moderate discrimination as their respective curves are neither very steep not very flat, and the overall measure largely targets persons at the average ability, slightly leaning towards below-average (with a mean *b* value of -0.21).

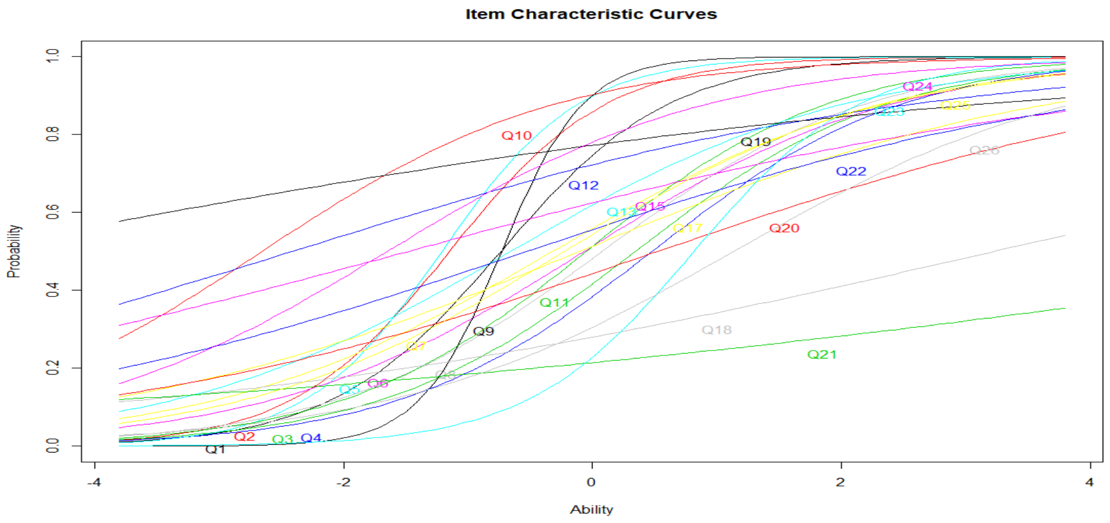


Figure 8: Item Characteristic Curves (ICCs)

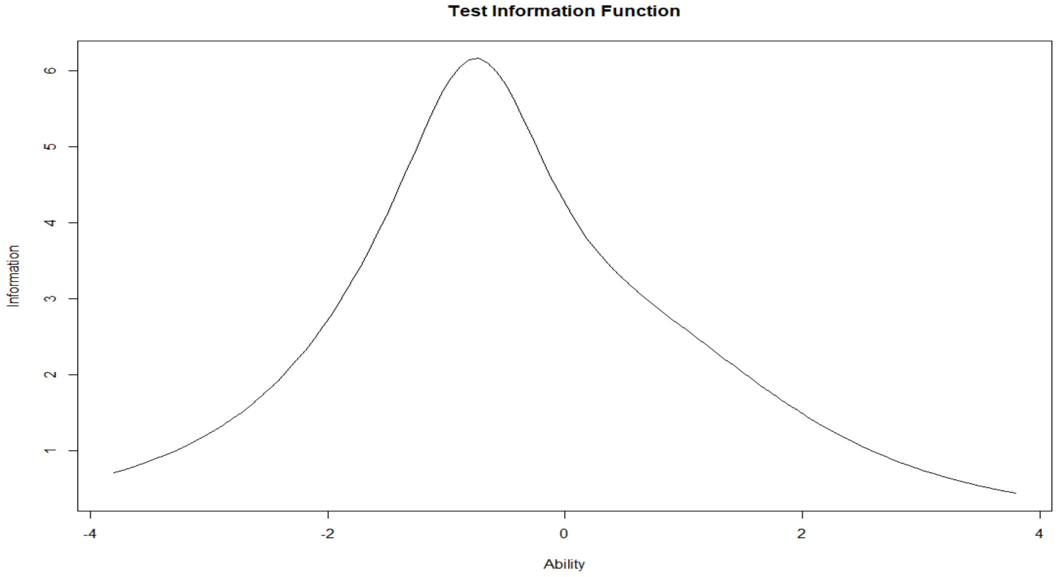
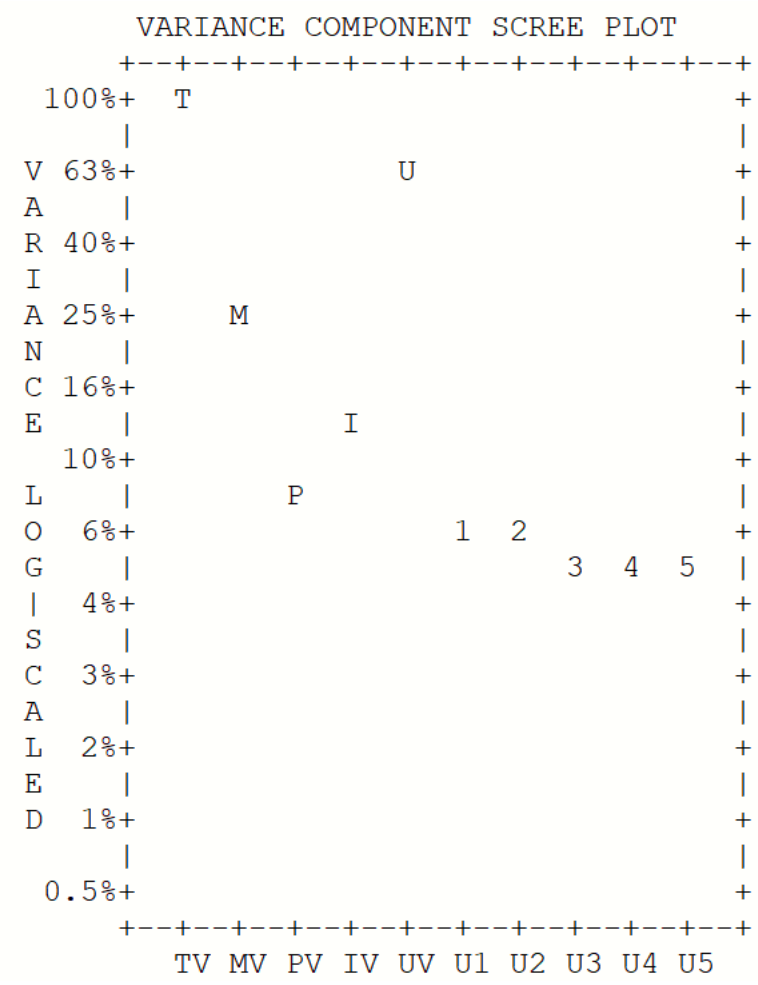


Figure 9: Test Information Function (2PL)

4.1.5 Rasch PCA

The final set of psychometric analyses performed with the purpose of exploring the psychometric properties of the measure was Rasch PCA. The total raw variance accounted for by the data had an Eigenvalue of 32.17, which corresponded to 25.4% of the total variance within the data. Total raw variance unexplained was 76.4%. Linacre (2012) informs that the amount of items with similar difficulty levels and the amount of persons with similar ability levels negatively correlate with the amount of variance explained by the assessment instrument. The percentages explained by persons and items were 9.7% and 15.7% respectively. Figure 10 presents a visual graphic of the results.



```
      VARIANCE  COMPONENT  SCREE  PLOT
     +--+--+--+--+--+--+--+--+--+--+--+
 100%+  T                             +
     |                                |
V 63%+              U                 +
A    |                                |
R 40%+                                +
I    |                                |
A 25%+     M                          +
N    |                                |
C 16%+                                +
E    |           I                    |
  10%+                                +
L    |        P                       |
O  6%+                 1  2           +
G    |                       3  4  5  |
|  4%+                                +
S    |                                |
C  3%+                                +
A    |                                |
L  2%+                                +
E    |                                |
D  1%+                                +
     |                                |
 0.5%+                                +
     +--+--+--+--+--+--+--+--+--+--+--+
        TV MV PV IV UV U1 U2 U3 U4 U5
```

Figure 10: Rasch PCA Variance and Components. Variance is represented by person (P), items (I), model (M), uniqueness (U), and total (T). The numbers represent possible components. Two components (1 and 2) were found to be significant enough to be taken into account.

Unexplained variance was reported using standardised residuals where the first construct had an Eigenvalue of 2.39 items (7.4% of the observed variance), and the second construct had an Eigenvalue of 2.14 items (6.6% of the observed variance). The other three contrasts had Eigenvalues below 2. Differences among items were investigated using the three item clusters reported. Pearson Correlations and Disattenuated Correlations are produced in the clustering of the items in which person abilities are measured with respect to each item cluster to be correlated with their measures from other cluster items, where Disattenuated Correlations use Pearson Correlations but ignore standard error (Linacre, 2012). Disattenuated Correlation was 0.06 for Clusters 1-3, 0.90 for Clusters 1-2, and 0.68 for Clusters 2-3. The closer to 1 the correlation, the more evidence that the clusters target the same latent variable, whereas correlations below 0.57 could be considered to address separate latent variables present (Linacre, 2012). The analysis suggests that the measure, in spite of having been developed with an aim to target nine standards as constructs, can be divided into two components, with the first component possibly having two sub-components.

In addition, a review of items according to their loadings showed that there were 13 items that were positively loaded and 11 items that were negatively loaded. The range for the positively loaded items was between 0.71 and 0.01 while the range for the negatively loaded items was between -0.02 and -0.58. Details of the items and their loadings are presented in Table 18.

Table 18: Modified CAK Item Loadings (N=74)

Item	Standard	Loading
Q5	Using Assessment Results for Decision Making	0.71
Q23	Using Assessment Results for Decision Making	0.45
Q6	Recognising Unethical Assessment Practices	0.42
Q8	Ensuring Quality Management of Assessment Practices	0.4
Q20	Developing Assessment Methods	0.4
Q24	Recognising Unethical Assessment Practices	0.26
Q17	Ensuring Quality Management of Assessment Practices	0.15*
Q4	Developing Valid Grading Procedures	0.1
Q18	Communicating Assessment Results	0.09
Q25	Keeping Accurate Records of Assessment Information	0.06
Q2	Developing Assessment Methods	0.03
Q15	Recognising Unethical Assessment Practices	0.01
Q22	Developing Valid Grading Procedures	0.01
Q21	Administering, Scoring, and Interpreting Assessment Results	-0.58**
Q12	Administering, Scoring, and Interpreting Assessment Results	-0.38
Q3	Administering, Scoring, and Interpreting Assessment Results	-0.36
Q13	Developing Valid Grading Procedures	-0.36

Table 18 (continued)

Q19	Choosing Appropriate Assessment Methods	-0.35
Q26	Ensuring Quality Management of Assessment Practices	-0.3
Q7	Keeping Accurate Records of Assessment Information	-0.23
Q10	Choosing Appropriate Assessment Methods	-0.21
Q11	Developing Assessment Methods	-0.11
Q1	Choosing Appropriate Assessment Methods	-0.06
Q9	Communicating Assessment Results	-0.02

* Signals the start of sub-component. ** Signals the start of the second component.

These results suggest that the modified CAK measure may not contain the modelled content domains or components; that is, the items on the measure may not represent the nine standards exactly the same way anticipated by the instrument. Instead, the remaining 24 items on the measure imply the presence of two components or constructs, with one of them having two sub-components. However, as unidimensionality vs. multidimensionality is not a dichotomous concept but more like a continuum or range (Linacre, 2012) and there was no sharp distinction between the variables on the two opposite edges, the results suggest that a unidimensional internal structure could be considered equally probable.

Additionally, a content analysis (to seek commonalities) of the scenarios and their corresponding items implied that apart from corresponding to separate latent variables, the components produced distinctive difficulty levels for the persons. In other words, an item difficulty analysis sorted by the components suggests that the participants exhibited dissimilar success rate for the two components. To elaborate on this finding, the content analysis suggests that the first component on the measure was related to more holistic topics or content areas in language assessment such as differentiating between assessment types based on purpose (i.e., formative vs. summative assessment), giving final grades to student performance, ethical considerations in assessment, and validity and reliability considerations in assessment. Even though they are closely related to each other (and thus described as sub-components rather than separate components), the items relating to the first sub-component (Component 1a) often addressed the topics of assessment type, and validity, the items relating to the second sub-component (Component 1b) targeted content areas like grading, and ethical considerations. Similarly, a search for common themes of the items in the second component (Component 2) suggested that the items in this component were often related to more specific topics and hands-on practices in assessment including choosing and developing appropriate tasks for immediate assessment, scoring student performance, and using, interpreting and communicating assessment results. Rasch model difficulty indicators (logits) were revisited to look for differences in participant performance in these components. The mean logit value was 0.03 for Component 1a, 0.17 for Component 1b (0.10 for overall Component 1), and -0.13 for Component 2, suggesting that the highest person achievement was in Component 2. Therefore, it could be possibly concluded that the participants had slightly more struggles in coping with the issues of differentiating between summative vs. formative assessment, grading, and ethical practices. However, it is worth noting that this kind of a generalization should be approached with caution, as it was shown earlier in the item analyses that the single most difficult item on the measure was about interpreting the interrelationship between test scores and percentiles.

4.1.6 Item Difficulties/Person Performance

The second research objective of the present study was to explore LAL of 4th grade pre-service EFL teachers at two university settings in Turkey at the knowledge base. The related research question was RQ(2): What is LAL knowledge base level of pre-service EFL teachers in two university contexts in Turkey? In order to inquire this question, data obtained using the modified CAK measure, whose psychometric properties were discussed in the previous pages, was used. 74 4th grade pre-service EFL teachers took the 27-question instrument, from which 3 items (Q14, Q16, and Q27) were removed due to poor psychometric properties, in separate proctored sessions. The fact that the data showed reasonable psychometric properties of the measure suggests that the responses given by the participants to the items could provide reliable insight into their language assessment knowledge base. The mean score of the participants was 13.5 out of 24 questions, denoting an average level of knowledge of the basics of AL. Table 19 presents the difficulty levels of each item as obtained from three different methodologies to item-person interactions (i.e., Rasch Analysis, CTT, and 2PL model), ranked according to difficulty from the easiest to the most difficult items on the measure.

Table 19: Items Ranked by Difficulty

Rasch Model		CTT		2 PL	
Item	**Difficulty (logits)**	**Item**	**Difficulty (Pcnt. Corrct)**	**Item**	**Parameter *b***
Q10	-2.11	Q10	0.89	Q19	-5.11
Q5	-1.46	Q5	0.82	Q10	-2.65
Q2	-1.26	Q2	0.8	Q12	-2.41
Q19	-1.07	Q19	0.77	Q24	-1.65
Q24	-0.98	Q24	0.76	Q15	-1.48
Q9	-0.9	Q9	0.74	Q5	-1.21
Q12	-0.66	Q12	0.7	Q2	-1.15
Q1	-0.58	Q1	0.69	Q1	-0.74
Q15	-0.23	Q15	0.62	Q9	-0.73
Q13	-0.16	Q13	0.61	Q13	-0.66
Q7	0.1	Q7	0.55	Q22	-0.53
Q22	0.1	Q22	0.55	Q7	-0.32
Q25	0.1	Q25	0.55	Q25	-0.23
Q17	0.23	Q17	0.53	Q17	-0.1
Q6	0.29	Q6	0.51	Q6	-0.07
Q11	0.29	Q11	0.51	Q11	-0.05
Q8	0.35	Q8	0.5	Q8	0.08
Q3	0.67	Q3	0.43	Q3	0.35
Q20	0.67	Q20	0.43	Q4	0.48
Q4	0.8	Q4	0.41	Q20	0.52
Q26	1.21	Q26	0.32	Q23	0.82
Q23	1.35	Q23	0.3	Q26	1.14
Q18	1.42	Q18	0.28	Q18	3.23
Q21	1.83	Q21	0.22	Q21	7.06

The ordering of the items follows exactly the same pattern for the Rasch and CTT models, while there are some slight discrepancies in the 2PL model, which could result from the difference in the theoretical approach to item-person interrelationship. In other words, even though the term Rasch model is often used interchangeably with 1PL model, this is called a mistake by Boone, Staver, and Yale (2013) because the two analytical families have differing philosophies in the sense that more parameters added in the IRT models to fit the data. It is also worth noting that person-ability estimates are made in the Rasch model without taking the distribution into account. Q10, for instance was, the easiest item according to Rasch and CTT models, while it was the second easiest in the 2PL model. Q21 was the most difficult item on the measure in all of the three models. Both Q10 and Q19 belong to Standard 1, which is related to choosing appropriate assessment methods, while Q21 belongs to Standard 3, which is about administering, scoring and interpreting assessment results. To further analyse the participants' achievement on the test with respect to the individual standards, item difficulty means were clustered into each corresponding standard (Table 20).

Table 20: Comparison of Participant Achievement by Standards

		Mean Difficulty					
		Rasch		**CTT**		**2PL**	
	Item	Logit	Mean	Pct Crrct	Mean	*b*	Mean
Standard	1	-0.58	-1.25	0.69	0.78	-0.74	-2.83
1	10	-2.11		0.89		-2.65	
	19	-1.07		0.77		-5.11	
Standard	2	-1.26	-0.10	0.8	0.58	-1.15	-0.23
2	11	0.29		0.51		-0.05	
	20	0.67		0.43		0.52	
Standard	3	0.67	0.61	0.43	0.45	0.35	1.67
3	12	-0.66		0.7		-2.41	
	21	1.83		0.22		7.06	
Standard	4	0.80	0.25	0.41	0.52	0.48	-0.24
4	13	-0.16		0.61		-0.66	
	22	0.10		0.55		-0.53	
Standard	5	-1.46	-0.05	0.82	0.56	-1.21	-0.20
5	23	1.35		0.3		0.82	
Standard	6	0.29	-0.31	0.51	0.63	-0.07	-1.07
6	15	-0.23		0.62		-1.48	
	24	-0.98		0.76		-1.65	

Table 20 (continued)

Standard	7	0.10	0.10	0.55	0.55	-0.32	-0.28
7	25	0.10		0.55		-0.23	
Standard	8	0.35	0.60	0.5	0.45	0.08	0.37
8	17	0.23		0.53		-0.1	
	26	1.21		0.32		1.14	
Standard	9	-0.90	0.26	0.74	0.51	-0.73	1.25
9	18	1.42		0.28		3.23	

Table 20 suggests that items measuring *Standard 1* (choosing appropriate assessment methods) were definitely the ones the participants were most successful in answering correctly. *Standard 1* was followed by *Standard 6* (recognizing unethical assessment practices) and *Standard 2* (developing assessment methods). On the other hand, the items found the most difficult by the participants on average belonged to *Standard 3* (administering, scoring and interpreting assessment results), *Standard 8* (ensuring quality management of assessment practices), and *Standard 9* (communicating assessment results). The rest of the standards produced relatively close-to-average difficulty values. An interesting finding regarding individual item difficulty levels is that Q21 and Q18, both of which, though falling under different standards, required the participants to do some mathematical reasoning.

4.1.7 Correlations

Lastly, in order to investigate the third research objective, regarding what factors (if any) affect LAL (RQ 3), several pieces of demographic information were collected from the participants before the administration of the measure. Such possible factors that were formulated into questions were identified through a review of the existing literature. The questions were related to their gender, current CGPA, their perceptions of preparedness for the EFL profession and for assessing students, and previous attendance to a workshop or seminar whose topic was specifically devoted to assessment. However, the categorical data provided by the participants was not large enough to create groupings to look for group differences (13 males – 59 females, and 5 previous attendances – 68 non-attendance assessment workshops or seminars. On the other hand, correlations were inquired using the ordinal and continuous data provided by the participants. First a scatter plot matrix

(Figure 11) was created to inspect the data visually. The plot hinted at a possible significant correlation between the ordinal variables of overall job preparedness and overall assessment preparedness, a significant correlation between the continuous variables of CGPA and the total score from the measure.

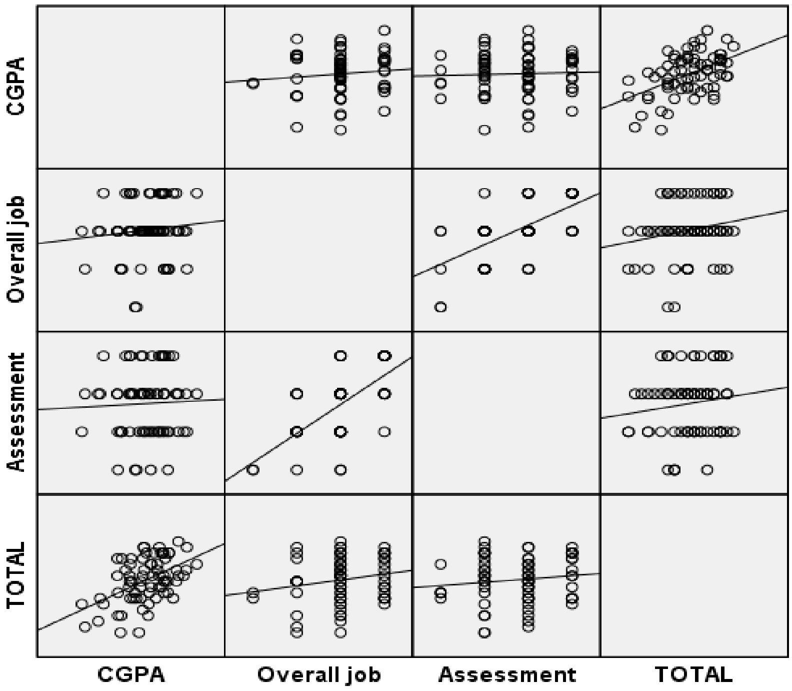


Figure 11: Scatter Plot Matrix

However, because there was no normal distribution in the overall job preparedness and overall assessment preparedness data, no further correlation analysis was performed between these two variables. The other possible correlation between the continuous variables of CGPA and total score was inquired (see Table 21 for descriptive statistics), and there was a positive but small correlation between the two variables (r = 0.52, n = 72, p = 0.01, R^2 = 0.27). The results of the Pearson correlation analysis suggest that even though there seems to be a positive relation between the participants’ overall success in school subjects and their AL, the importance of this relationship, due to a small effect size, is rather small and cannot explain more than around 30% of the variance on the scores from the measure.

Table 21: Descriptive Statistics for CGPA and Total Score

	CGPA (out of 4.00)	Total Score
Number	72	74
Minimum	2.25	1
Maximum	4.00	21
Mean	3.22	14.07
Std. Deviation	0.38	4.27
Skewness	-0.57	-0.62
Kurtosis	0.26	0.10

4.2 Discussion

A discussion of the findings is presented in the following paragraphs in relation to each research question of the study.

4.2.1 Research Question 1

The primary research objective of the current study was to find out about the psychometric properties of the modified CAK. As discussed in the rationale section in Chapter 1, there is a need for objective and accurate assessment tools to asses EFL teachers' AL due to the gaps experienced by EFL teachers with respect to assessment knowledge and practice. Addressing this need requires an accurate, valid and reliable instrument to assess EFL teachers' LAL. There have been a number of such instruments designed and developed to this day in the literature. It is possible to roughly divide these instruments into two broad categories with respect to the approach they adopt towards assessing AL: (a) those using a survey approach where participants are surveyed to elicit information about a number of issues such as how they feel about the training they received on assessment, and how they feel about their strengths and weaknesses, and (b) those adopting a more objective approach where participants are asked to answer questions and are tested for their knowledge of assessment. The survey-type measures are often used to inform researchers and policy makers on the strengths and weaknesses of teachers and their training needs as well as any possible need to improve training programs. Vogt and Tsagari's (2014) Teachers' Questionnaire is a popular example of this type of measures. Measures that test teachers' knowledge of AL could be exemplified by King's (2010) Criterion-referenced Assessment Literacy, Plake's (1993) TALQ and

Mertler's (2003) CALI. The present study adopted the second approach in choosing an instrument to assess pre-service EFL teachers' LAL. The literature presents a number of objective measures of assessing teacher AL at the knowledge base (refer to Chapter 2 for details). The first stage in choosing one to use in this study involved the comparison of these measures in terms of their theoretical ground. One of the most frequently used theoretical frameworks for such measures is the *Standards for Teacher Competence in Educational Assessment* (AFT, NCME, & NEA, 1990). TALQ, ALI and CALI are the most prominent assessment instruments based on the *Standards* document. A number of research studies have so far employed, adapted and modified versions of these measures, and produced high levels of reliability estimates (e.g., Alkharusi et al., 2011; Chapman, 2008; Mertler, 2003; Mertler & Campbell, 2005; Plake, 1993). Because the current study focussed on EFL teachers, an assessment instrument designed by Tao (2014) adapting from Mertler and Campbell's ALI (2005) measure was used in this study. Called CAK, the measure consists of 27 items, 11 of which were adapted from ALI while the other items were developed by the author based on the *Standards*. Following an extensive literature review, the author included two more standards to enhance the content validity of the measure in addition to the seven original standards. The measure was converted into a more user-friendly format, attaching all of the questions to three scenarios in which classroom language teachers need to make assessment-related decisions. There are three scenarios in total, each with their corresponding 9 items. Each item in each scenario addresses corresponding outcomes defined by each standard. The scenarios and the items were tailored for an EFL classroom context. The measure was modified for use in this study, as well. The context was adapted for an EFL classroom setting in Turkey. A number of format, content and wording alterations were also applied by the researcher to address a number of possible problems with the items, anticipated by a panel of assessment and language assessment specialists. One item was completely rewritten. A total of 74 4th grade pre-service EFL teachers completed the measure.

Following the administration of the measure, a number of psychometric analyses were performed with the data provided by the participants. These item and test analyses consisted of a 1PL model, CTT methodology, 2PL model, and Rasch PCA.

The first analysis (1PL) was carried out to make initial sense of the information provided by the data. Firstly, the data was tested for model fit; in other words, for whether the data is suitable for 1PL model. Following the establishment of the model fit, further analyses were made about the items and the overall measure. The initial analysis showed that the ability range tested by the items on the measure was between -3.15 and 2.47. In the model, 0 refers to the average ability, and any value below 0 refers to below-average ability while any value above 0 denotes above-average ability. 11 items on the measure matched with above-average ability while 16 items addressed below-average ability. The average difficulty of all items was found to be -0.20, meaning that the measure had a moderate difficulty level. A visual analysis of ICCs showed that the items are normally dispersed along the 0 – 1.0 probability line. Similarly, the visual analysis of IICs revealed that all of the items on the measure formed three clusters below, at and above the 0 ability level, suggesting that the measure roughly contained three groups of items: easy items, moderate items and difficult items. Test Information Curve also suggested a similar conclusion, with its bell curved shape. In other words, it demonstrated that the measure, as was expected, provided the greatest amount of information around the average ability. Test information function of the measure indicated that the measure was able to effectively account for approximately 83% of the participants' latent trait. These findings were considered to be an initial evidence of acceptable psychometric properties of the measure in terms of item and measure difficulty levels and internal consistency.

However, a highly noticeable observation was made about the item labelled Q10. It appeared to be positioned rather farther from other items on the probability line, suggesting it was much easier compared to the other items. This item (refer to Appendix C) was affiliated to the standard related to choosing appropriate assessment methods (Standard 1), but it had also some implicit reference to validity. 66 participants (89%) got this item correct, a facility level 33% greater than the mean difficulty level. Because the item was found to be too easy by almost all persons, it produced a rather low discrimination index (0.17). So, any possible replication of this item should take caution, as these results might have been caused by either successful learning or poor psychometric properties. Possible psychometric problems associated with this item are likely to stem from poor or weak creation or organization of the options. The options analysis indicated that 4 participants selected Option A, 3 participants selected option E, and only one participant selected Option C. Therefore, it should be noted that any replication or recreation of this item should be done with more plausible options.

After the initial 1PL model, CTT methodology was used to examine the test statistics including descriptive statistics, item statistics (i.e., difficulty and discrimination), and options statistics. Concluding from the test statistics, it could be argued that the measure produced a Cronbach's alpha value of 0.73, which could be considered a good value for the sample size. Three participants with the highest number of correct answers got 20 items right out of 24 items. The mean number of correct answers was 14, and the minimum was 1, with a standard deviation of 4.10, which suggests that there is much variability in the test scores. The mean for item difficulty was 0.53, which indicates that the measure had an average difficulty level for the sample. With the faulty items eliminated, the mean discriminating index of the measure was 0.40, meaning that the overall measure is able to effectively discriminate high-achieving persons from low-achieving persons. There is a general tendency in the literature to accept a discrimination index value between 0.10 and 0.30 to be fair, and a discrimination index value of 0.30 and more to be good. Even though the vast majority of the items (19 items) had values greater than 0.39, three items (Q9, Q1, Q13) had by far the highest discrimination values with 0.70, 0.61 and 0.60, respectively. Q9 was affiliated to the Standard related to communicating assessment results (Standard 9). In Q9, the participants were asked to choose what a classroom teacher should do when explaining to the students the basis for assigning course grades. Q1 (Standard 1) required the participants to know about portfolios as assessment tools to monitor student performance over time. And Q13 (Standard 4) required the participants to be able to differentiate between fair and unfair criticisms to an assessment decision taken by a classroom teacher in which the teacher used only one assessment instrument to give final grades to the students. One of the items (Q21) on the measure had a discriminating index value far below the average compared to other items (0.19). Q21, which was affiliated to the standard related to score interpretation (Standard 7), required the participants to make sense of and interpret the relationship between mean scores and standard deviation values of an assessment instrument. This item was also the most difficult item on the measure with a difficulty level of 0.22, answered accurately by only 16 participants. The obvious cause for low discriminating power for this item appears to originate from the fact that even the high-performing participants got this item wrong. The number of persons in the high-achieving group who got this item wrong (21) is greater than the number of persons in the same group who got this item right (16). A similar observation was reported for Q18, the second most difficult item on the measure, as well. Even though Q18 was affiliated with the standard related to communication of assessment results (Standard 9), it involved a score interpretation process similar to the one in Q21. The discrimination index for this item was good (0.33), but the distribution of high-performing persons across the options was equal for the key and one of the distractors. These observations suggest that most of the participants, even those in the high-achieving group, had a considerable struggle on items that assess their ability to understand and interpret scores mathematically.

Three items were reported to have very low discrimination indices (Q14, Q16, and Q27). These items were associated with the standard related to using assessment results for decision-making, the standard about keeping accurate records of assessment information, and communicating assessment results respectively (Standard 5, Standard 7, and Standard 9). Discrimination indices for these items were 0.05, 0.08, and -0.2. According to the options analysis of Q14, even though the key received more answers compared to the distractors, an equal number of 7 high-performing persons selected the key and one of the distractors, which led the item to fail to discriminate between groups. A closer review of the item's content indicated indeed problems in the item. The item aims to test the ability to differentiate between norm-referenced and criterion-referenced assessment information; however, the way it was formulated and worded appears to cause the item to end up having three correct answers. Similarly, in Q16, which asked the participants to distinguish between the purposes of keeping assessment records, there was no significant discrimination because more high-achieving persons opted for two distractors rather than the key. The content review also found that there might be no correct answer at all in the item. And Q27, which required the participants to establish and understand the relationship between raw scores and percentile ranks, had two distractors that received more answers than the key from the high-achieving group of persons. It produced a negative discrimination value. A content examination of the item actually found no problems with the content, but it could have been the way of formulation and wording of the item that created the problem. Therefore, these three items were removed from the measure, and the following psychometric analyses did not take these items into consideration.

The next stage of psychometric investigation was performed with Rasch analysis, this time using logit values and item-person maps. The first observation was regarding the number of iterations attempted by the analysis, which was four, suggesting that the algorithm did not have difficulty figuring out the parameters of item difficulty and person ability. Reliability and separation statistics for items and persons were also checked, and it was found that reliability values were 0.72 for persons, and 0.91 for items, both of which could be considered to be high reliability figures, suggesting the measure could be replicated and could produce similar results. And separation values were 1.61 for persons and 3.25 for items, suggesting that the measure is capable of differentiating between around one and a half groups, and that there are slightly more than 3 groups of item difficulty. Interpretation of person separation requires some caution, as it is possible that this could have occurred due to small sample size, and because the measure was taken by a highly homogenous group of people (i.e., 4th grade pre-service EFL teachers). In other words, judging from the separation statistics concerning the items, the measure could produce higher separation values for a more heterogeneous group such as in-service EFL teachers with different educational backgrounds, experience levels, and types of schools they work for, etc.

A visual analysis of the measure through item-person maps (Wright maps) provided initial insight into the results of the analysis. A Wright map presents the range of item difficulties and person abilities on the same scale using the same logit values. The visual analysis found that the bulk of the items and the bulk of the persons were located around slightly above the mean on the scale, suggesting a normal distribution for both items and persons. There were 27 persons one standard deviation above the mean and 25 persons one standard deviation below the mean. There was no person two standard deviations above the mean, while there was only one person two standard deviations below the mean. That person got only one item correct in the entire measure. Similarly, there were 10 items one standard deviation above the mean and 6 items one standard deviation below the mean. No items fell outside the two standard deviations range above the mean; however, there was an item (Q10) located around more than two standard deviations below the mean. The same item was already reported in the 1PL and CTT item analyses to have been by far the easiest item on the measure. The visual analysis of the Wright map found no significant gaps (except for the mentioned outliers) between the clusters of items and persons. In other words, there was a certain number of items matching the corresponding ability levels of almost all persons (refer to Figure 6 in Chapter 4 for details about item-person matchings).

The test statistics obtained from Rasch analysis were also inspected, and the ranking of individual items based on their difficulties was found to follow the same pattern as the results produced by CTT methodology. Q10 was the easiest item with a logit value of -2.11 and Q21 was the most difficult item with a logit value of 1.83. Standard error of measurement was also examined for individual items and for the overall measure. The range for the error values for individual items was between 0.25 and 0.40 (with a mean of 0.28), which reveals high confidence in the measurement. Mean logit value was 0.01, suggesting that the measure can overall be said to address average ability. The infit and outfit statistics provided by Rasch analysis are good indicators of possible misfits among the variables. In other words, they help understand the fit relationship between the model and the data. Although the literature suggests that infit and outfit values between 0.5 and 1.5 are productive for measurement (Linacre, 2012), the general tendency is to accept the range of 0.8 and 1.2 as the most productive. The infit values of the items on the measure ranged from 0.74 to 1.14 (mean: 1.0), and outfit values ranged from 0.60 to 1.29 (mean: 0.97), suggesting that there was no item that could be considered an underfit or and overfit.

To check and verify the findings from the previous item and test analyses, 2PL model was used. The theoretical background of this model in relation to the concept of latent trait is very similar to Rasch model, but it takes into account the parameter of item discrimination. This set of analysis produced mostly similar results to Rasch analysis and CTT methodology for item and test analyses. Having established that the data fit the model, the data were firstly visually inspected. The positioning of ICCs on the ability-probability scale was in line with what was expected taking the previous analyses into consideration. However, an inspection of

IICs revealed that the majority of items contained moderate amounts of information individually, but the IIC of Q9 stood out quite remarkably compared to the other items. Q9, which was also the item with the highest discrimination power (*a*: 3.02), contained the largest amount of information. A closer inspection of the ICC belonging to Q9 found that the curve had a close-to-perfect "S" shape, where persons with an ability value between -4 and -1 had almost 0 probability of answering the item correctly, while the probability started to increase to 0.5 at around -1.0 ability and move towards 1.0 at around 0 ability, suggesting that the item was successful in discriminating between average ability persons and below average ability persons. A visual review of Test Information Function found that the measure provides the largest amount of information for persons between the ability range of -1.0 and 1.0 (i.e., average ability), denoting normal distribution once more.

Another item with a remarkably high discriminating value was Q6 with an *a* value close to 2.0. This item, associated with the standard related to using assessment results for decision making (Standard 5), had a probability of correct answering at the ability level starting with 2.0, and around 0.8 probability at 0 (average) ability, while having less than 0.2 probability at -2.0 ability. It shows that this item was able to effectively discriminate above-average, average and below-average persons from each other. However, the overall discrimination analysis of the items within this model showed a moderate success in discriminating groups of persons with differing abilities (mean *a*: 0.87). Three items (Q18, Q19, and Q21) on the measure produced discrimination values that could be considered very low according to the recommendation made by Baker (1985). Due to the interaction between item difficulty and item discrimination, it could be argued that these very low discrimination values were because of the fact that these items were too easy or too difficult for the participants (refer to Table 17 in Chapter 4 for *b* and *a* parameter values). Figure 12 presents the ICCs for these items.

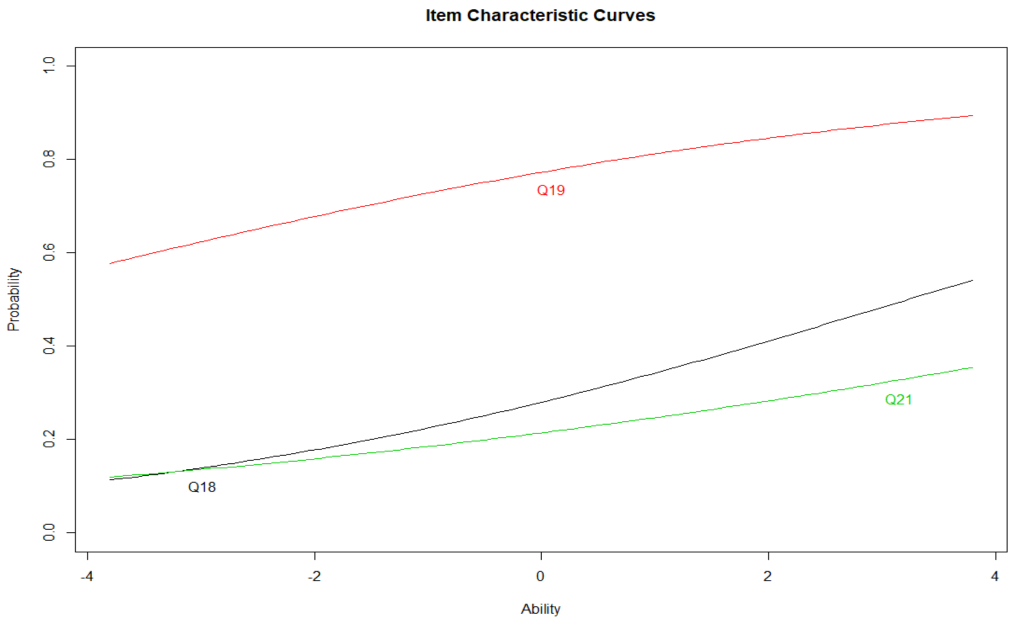


Figure 12: ICCs Belonging to Q18, Q19, and Q21

As can be understood from Figure 12, ICCs of these three items formed straight and linear lines, being very far from representing an “S” shape, probably due to their too low (Q19) or too high (Q18 and Q21) difficulty values. These items denote very similar probability values for persons at all ability levels. Q18 (Standard 9), which was found very difficult by the participants, required them to show their knowledge of understanding the concept of percentile. Q19 (Standard 1), on the other hand, was the easiest item according to the 2PL analysis results. It was likely to be answered correctly (more than 0.5 probability even for persons at around -4.0 ability level) by all persons, therefore providing very little amount of information. This item asked the participants to choose the correct type of assessment relevant to a situation. Q21 (Standard 3) required the participants to establish and interpret the link between test scores and standard deviation values.

It could be argued that the results of the analytical techniques so far discussed (1PL model, Rasch analysis, CTT methodology, and 2PL model) produced comparable pieces of insight into the psychometric properties of the modified CAK measure and the items it contains. Although some of the 2PL findings had minimal departures from the other techniques as to item discrimination and item difficulty, all four models could be said to find comparable results, leading up to similar conclusions about the psychometric properties of the modified CAK measure. Therefore, it is worth noting that this finding comparing the techniques resonates with the findings of Fan (1998) who argued that the person and item statistics produced by CTT approaches and IRT approaches are “highly comparable” (p. 14).

The final analytical technique used to explore the psychometric properties of the modified CAK was Rasch PCA, which aimed to investigate the dimensionality within the measure. The analysis found that the raw variance explained by the measure was 25.4%, while total raw unexplained variance was 74.6%, which suggested that there could be some components present within the measure accounting for a certain extent of the unexplained variance (refer to Chapter 4 for details). Findings from the analysis implied the presence of two components within the measure even though the measure was originally designed to address nine different constructs (i.e., the nine standards for teacher competence in educational assessment of students. The first component appeared to have two sub-components. Table 22 presents the possible components of the measure and the items that seem to be belonging to these components.

Table 22: Clustering of Items and Standards into Two Components

	Item	Standard
Component 1a	Q5	5
	Q6	6
	Q8	8
	Q20	2
	Q23	5
	Q24	6
Component 1b	Q2	2
	Q4	4
	Q15	6
	Q17	8
	Q18	9
	Q22	4
	Q25	7

Table 22 (continued)

Component 2	Q1	1
	Q3	3
	Q7	7
	Q9	9
	Q10	1
	Q11	2
	Q12	3
	Q13	4
	Q19	1
	Q21	3
	Q26	8

According to Table 22, Standards 2 and 6 were completely clustered into Component 1, and Standards 1, 2, and 5 were exclusively clustered into Component 2, while the rest of the standards had items that belonged to either Component 1 or Component 2. A review of the content areas of the items clustered into the components revealed that Component 1 was mostly related to assessment topics and concepts of assessment types, validity, grading, and ethical considerations, while the items in Component 2 were largely related to more specific topics and hands-on practice-related issues such as choosing and developing tasks for immediate assessment of students, scoring student performance, and using, interpreting and communicating assessment results. Reduction of the nine constructs (Standards) into two components resonates with the findings of Ryan (2018), who investigated the psychometric properties of a modified version of Mertler's (2003) CALI instrument, which was based on the same *Standards* document designed by AFT, NCME, & NEA, 1990), which is indeed understandable given that the skills and outcomes described by the standards are closely related to and interact with each other. In other words, not only are they the building blocks of AL, which together constitute the knowledge base of AL, but understanding of one also helps understand the other standards and influence decisions regarding them.

The psychometric properties of the modified CAK demonstrated its effectiveness as a potential measure to be used in the assessment of EFL teachers' LAL knowledge base, a finding similar to that of Mertler and Campbell (2005) who found CALI’s overall reliability coefficient to be 0.74. Through a review of individual item characteristics, they also found the original ALI measure items to have satisfactory item discrimination values. 25 (71%) items out of 35 had a discrimination index above 0.30 while 20 (74%) items out of 27 on CAK had items falling in this range. Furthermore, these results were also confirmed through the other psychometric analyses (IRT models and Rasch model), in addition to CTT methodology. This study also found that the measure could have two principal components, resonating with Ryan’s (2018) finding, who adapted CALI (also based on the *Standards* document), which means that despite being interrelated, these two components might be addressing different latent traits of AL.

The results of the psychometric analyses of the CAK instrument suggest that the instrument can be used to generate evidence-based information about the assessment knowledge of EFL teachers in Turkey with high confidence in terms of validity and reliability. The information obtained can be used to assess and evaluate EFL teachers’ AL-related strengths and weaknesses as well as their training needs in order to help shape any possible educational and training activities. Such information may be useful at both pre-service and in-service level. However, it is worth noting that any replication of the instrument or the recreation of the items in the instrument should take the needs of the relevant context and purposes into consideration. In other words, it would be safe to highlight that the specifications provided by the *Standards* document, with local needs and conditions taken into account, could constitute a robust framework in the assessment of AL of EFL teachers.

4.2.2 Research Question 2

The second research objective of the study was to find out about LAL of pre-service EFL teachers in two higher education contexts in Turkey. The modified CAK was used in the present study to assess pre-service EFL teachers’ LAL at the knowledge base. Even though a complete understanding of teachers’ AL is only possible through a comprehensive evaluation including observation of teachers when performing assessment practices in order to have an understanding of their knowledge, skills and ability of putting their knowledge into practice, research suggests that AL starts with the knowledge base, which significantly correlates with the success of practice (Xu & Brown, 2017).

The data collected from the participants were analysed through the perspectives of several different theoretical frameworks, i.e., 1PL IRT model, Rasch model, traditional CTT model and 2PL IRT model, all of which produced similar results regarding both individual items and the overall measure statistics. The mean raw score out of 24 was 13.5, slightly above the midpoint. This translated into an average measure difficulty (56%). This finding suggests that the performance of the participants on the measure was average, meaning that an average participant is able to answer approximately half of the items correctly. In other words, pre-service EFL teachers in the study had knowledge of around half of the content areas covered by the Standards document. This finding is consistent with the findings of previous research (Fulcher, 2012; Inbar-Lourie, 2008, 2013; Lam & 2015, Taylor, 2013; Tsagari & Vogt, 2017; Volante & Fazio, 2007). No participant was able to get all the items right. There were three top-scoring participants, all with the same raw score of 20. A common item answered inaccurately by these three top-scorers was Q21, which was also found to be the most difficult item on the entire measure. This item was associated with the standards related to administering, scoring, and interpreting assessment results (Standard 3). Similarly, Q3 (Standard 3) was answered inaccurately by two of these three participants.

Q10 (Standard 1) was remarkably the easiest item according to the results of Rasch model and CTT model. The vast majority of all the participants answered this item correctly, suggesting that they know that a teacher who wishes to assess his/her students' understanding of a subject after teaching the subject should design an assessment instrument whose items are consistent with the content and skills specified in the course learning outcomes. Next group of items with considerably lower difficulty levels compared to the other items were Q5, Q2, Q19, Q24, and Q9. These were items assessing knowledge of summative vs. formative assessment, assessment of writing performance, authentic assessment, unethical assessment practices, and assigning grades. On the other hand, Q21 was found to be by far the most difficult item on the measure with a considerably higher difficulty value compared to the other items. This item required the participants to do a certain degree of mathematical reasoning to interpret student performance. The scenario presented them with a situation in which a student wanted to compare his performance on tests of reading and writing, where the student's scores, mean scores and standard deviations for each test were given. The next cluster of difficult items with considerably higher difficulties than the other items included Q18, Q23, and Q26. Q18 required the participants to make sense of the concept of percentile, while Q23 tested their knowledge of the concept of summative assessment, and Q26 assessed their understanding of the role of practicality as compared to the other fundamental principles of assessment.

A standard-wise comparison found that the Standard 1 (choosing appropriate assessment methods) was clearly the standard the participants found the easiest to cope with. It was followed by Standard 6 (recognising unethical assessment practices). In contrast, Standard 3 (administering, scoring, and interpreting assessment results) and Standard 8 (ensuring quality management of assessment practices) were the two most challenging standards for the participants. Table 23 presents the ranking of the standards according to their difficulties from the easiest to the most difficult.

Table 23: Ordering of Standards by Difficulty

Rank	Standard	Logits
1	1	-1.25
2	6	-0.31
3	2	-0.1
4	5	-0.05
5	7	0.1
6	4	0.25
7	9	0.26
8	8	0.6
9	3	0.61

It is also worth looking at participant performance according to the components found within the measure following the Rasch CPA. Even though one of the components appeared to possibly contain two sub-components, the analysis found two possible components present. The first component was found to largely involve content areas in assessment such as assessment types (differentiating between dichotomies), giving course grades, ethical considerations in assessment, and validity and reliability considerations in assessment. The second component mostly involved content areas such as choosing and developing appropriate tasks for immediate assessment, scoring student performance, and using, interpreting and communicating assessment results. Despite not having remarkably different difficulty values, it could be argued that the participants were slightly more successful in the second component.

The findings from all these reviews suggest that the sample in this study exhibited an average level of success in terms of LAL, which was measured against the *Standards* document. They demonstrated no serious problems in terms of choosing appropriate assessment methods for the assessment of student success. They were also relatively more successful in recognising unethical assessment practices. On the other hand, they had serious difficulties particularly when it came to mathematically calculating, making sense of, and interpreting test scores and evaluating student performance using those raw test scores and other test statistics such as percentiles and standard deviations. They also found it relatively more challenging to cope with content areas regarding enhancing the quality management of assessment practices. All in all, there were few standards that the participants exhibited an excellent or a very poor performance. The majority of the standards, and thus the majority of the items, received an average performance from the participants.

However, given that the participants, 4th grade pre-service EFL teachers having completed the majority of the course load, had just taken a course on assessment at the time of data collection, the performance can be considered to be lower than expected, which implies a serious lack of assessment knowledge. This finding is similar to what has been found regarding EFL teachers' LAL both around the world and in Turkey (Hatipoğlu, 2015; Lam, 2015; Öz & Atay, 2017; Tsagari & Vogt, 2017; Volante & Fazio, 2007; Xu & Brown, 2017). In other words, knowledge of classroom assessment practices and activities is limited, and there is a need for more and quality training including both theoretical and practical aspects of language assessment accompanied with content from measurement component of educational assessment.

These results can be interpreted to be pointing to some alarming and worrying conclusions. Although there is a need for a more comprehensive consideration of the Turkish pre-service EFL teachers' LAL in a way to incorporate their both theoretical knowledge of and practical skills related to assessment activities and practices, including language-specific assessment elements, the lower-than-expected performance of the participants on the measure requires further research on the causes of this low performance and possible suggestions about what can be done to change the status quo. In fact, in her comprehensive research investigating the English Language Testing and Evaluation (ELTE) course in the English Language Teacher Education programs in Turkey through the perspectives of pre-service EFL teachers and ELTE instructors, Şahin (2019) endeavoured to carry out a detailed analysis of the ELTE course and how it contributes to the development of LAL of pre-service EFL teachers. The findings of the present study can be considered to be similar to those found by Şahin, who concludes that pre-service EFL teachers in Turkey complete the English Language Teacher Education program without attaining adequate theoretical knowledge and practical skills needed in the assessment and evaluation of English language learners. The problem is partly caused by the inadequate amount and content diversity of the assessment courses and by ELTEC instructors not employing a holistic approach in the teaching of assessment topics. One particular finding from this study was that measurement concepts, especially those requiring mathematical and statistical reasoning and understanding, appear to be paid little focus. Therefore, it is worth noting that more similar research is needed to help better understand why pre-service EFL teachers in Turkey lack the expected assessment knowledge and skills in addition to how to tackle this problem with particular focus on the English Language Teacher Education programs, educational policies, curricula and course syllabuses to understand what is going on in the classroom during the ELTEC instruction. It is quite significant that EFL teachers graduating from their programmes receive adequate and quality training in assessment to apply the theoretical knowledge of the concepts in assessment into their practices in an effective and successful way.

4.2.3 Research Question 3

The third research objective of this study was to explore what factors, or background characteristics, if any, contribute to the development of LAL of EFL teachers. A large of number of research studies in the literature endeavoured to find out about the AL literacy. Such studies often focussed on in-service teachers, and their research design was often tailored to elicit information about teachers' assessment needs. For instance, King (2010) and Mertler (1999) found that the quality of assessment training teachers underwent impacted their knowledge base of AL. Another background characteristic with possible effect on the development of AL was found to be teaching experience (Alkharusi, 2011; Chapman 2008). Though with mixed results, academic qualification was counted as another characteristic that could be influencing success in assessment (Chapman, 2008; King, 2010). Gender was also explored as a potential factor by Alkharusi (2011), who found that teachers' self-perceived assessment competence significantly differed regarding gender, and that female teachers considered themselves to be more competent in item writing and communicating assessment results to stakeholders. A large body of research has found professional development as an important factor enhancing teachers' AL (Mertler, 2009). In other words, teachers in schools that provided in-service training to teachers on AL had better performance in assessment practices. Teaching hours, class size, and assessment experience as students were also reported to have been found as factors possibly contributing to the development of AL (Tao, 2014). In summary, the possible factors to contribute to AL are pre-service assessment training, teacher experience, academic qualification, gender, professional development, teaching hours, class size, and assessment experience as students.

Because the present study specifically focussed on pre-service EFL teachers in assessing their knowledge base of language assessment literacy, the sample constituted a rather homogenous structure, where it was not possible to collect a large number of categories of background information from the participants. Of all the background categorical information collected from the participants, the only statistically significant correlations were between their self-reported perceptions of overall job preparedness and assessment preparedness, and between their CGPA and total scores from the modified CAK. The participants reported a lower rating (2.7, below "prepared") for their preparedness for assessment than their rating for overall job preparedness (3, "prepared"), which clearly indicated a need for some intervention, possibly through a revision of the training program. One interesting finding was regarding the positive correlation between these self-reported ratings. The more prepared the participants felt about their preparedness for the overall teaching job, the more confident they felt about their preparedness for student assessment, which is also an indication of the close relationship between instruction and assessment (Malone, 2013; Popham, 2009; Stiggins, 1999; White, 2009). In addition, the significant positive correlation between CGPA and modified CAK total score suggests that teachers who are more successful in EFL subjects in general are more likely to be successful assessors conducting high quality assessment practices, a finding supporting Chapman (2008) and (King) 2010. Therefore, it could be argued that that by shifting up the emphasis of assessment within the teacher training programs, teachers could be aided to become more assessment literate educators.

Due to the methodological limitations, the present study was able to explore only few of the possible factors influencing the development LAL of pre-service EFL teachers. However, closely related to RQ2, RQ3 requires an extensive and separate inquiry into the possible factors contributing to the development or absence of the required LAL skills of EFL teachers in Turkey. Gaining insight into this research question may also help better understand the inadequate LAL levels experienced by pre-service EFL teachers. The literature has so far counted pre-service assessment training, teacher experience, academic qualification, gender, professional development, teaching hours, class size, and assessment experience as students as possible factors. Each of these and possible new factors should be investigated thoroughly through extensive research that may employ a review of a number of components including teacher education policies, English Language Teacher Education programs, institutional approaches to the subject of assessment, related curricula and syllabuses, the quality of the instruction of the language assessment courses, and EFL teachers’ perceptions of and perspectives on language assessment.

CHAPTER 5

CONCLUSION

This chapter consists of three sections. Section 5.1 outlines a review of the research objectives of the study and the procedures followed in the investigation of the objectives. This is followed by a discussion of the findings of the study in Section 5.2, elaborating on the implications and conclusions from the findings. Finally, Section 5.3 discusses the limitations of the study and future research directions.

5.1 Overview of the Study

5.1.1 Overview of Rationale of the Study

The rationale of the study was established taking into consideration the existing literature which argues that even though assessment is a crucial part of instruction and education in general, and there are numerous benefits of high quality assessment with respect to student learning and teacher instruction, a great majority of teachers at all levels exhibit low levels of AL. Therefore, the study's main research objective was to determine whether a modified LAL measure based on the *Standards* set by AFT, NCME and NEA (1990) could possibly be used in the assessment of EFL teachers' LAL. The primary purpose was to investigate the psychometric properties of the measure to obtain insight into the measure's validity and reliability so that such insight could be used to reach a decision on whether such a measure based on the standards as constructs might be employed to learn about EFL teachers' LAL at the knowledge base, which, in turn, may feed educational policies. A secondary purpose derived from the main purpose was the possibility of using the measure to understand the current language assessment literacy of EFL teachers. In addition, the study also aimed to explore whether there were any factors impacting EFL teachers' LAL. Three research questions that emerged from the rationale were as follows:

1. What are the psychometric properties of the adapted Classroom Assessment Knowledge instrument, devised to assess EFL teachers' language assessment literacy knowledge base?
2. What is the language assessment literacy knowledge base level of pre-service EFL teachers in the higher education context in Turkey?
3. What factors, if any, affect language assessment literacy of pre-service EFL teachers in the higher education context in Turkey?

5.1.2 Overview of Methodology

The study adopted a quantitative research design in order to explore the issues formulated by the research questions. It aimed to (a) examine the psychometric properties of a LAL measure, (b) find out about EFL teachers' LAL, and (c) explore the relationship between language assessment literacy scores and demographic variables provided by the participants.

5.1.2.1 Instrument (CAK)

A measure to test EFL teachers' language assessment knowledge was employed in the study. The measure was modified and adapted from the CAK measure developed by Tao (2014). The original CAK used the Standards for Teacher Competence in Educational Assessment of Students (AFT, NCME, & NEA, 1990). It contained 27 items in total, 10 of which were adapted from the ALI instrument (Mertler & Campbell, 2005). The rest of the items in the original measure were developed by the author based on an extensive search of the literature. In order to achieve integrity and set the context, the measure presented three scenarios, each having 9 items. In the modified version of CAK used by this study, a number of alterations were made in the content of both scenarios and items following a review of the measure by a panel of assessment specialists and language assessment specialists. The measure expanded on the seven standards by adding two more standards. The two additional standards aimed to address the criticisms to the original standards over their narrow aspects. It was considered that all key stages of the assessment process were covered by the nine standards (refer to Chapter 2 for details). The nine standards addressed by the modified measure included:

1. Choosing Appropriate Assessment Methods
2. Developing Assessment Methods
3. Administering, Scoring, and Interpreting Assessment Results

4. Developing Valid Grading Procedures
5. Using Assessment Results for Decision Making
6. Recognising Unethical Assessment Practices
7. Keeping Accurate Records of Assessment Information
8. Ensuring Quality Management of Assessment Practices
9. Communicating Assessment Results

There were three questions associated with each standard, and they followed the same order in each contextualised scenario. In other words, each scenario tested the nine standards with the corresponding nine items in the same order.

5.1.2.2 Data Collection

The research employed convenience sampling procedures to select the participants of the study. 4th grade pre-service EFL teachers from two prominent state universities in Ankara participated in the study to complete the measure. All of the participants had taken at least one course on English Language Testing by the time they took part in the study. A total of 74 participants completed the measure on separate proctored sessions. The participants were given approximately 45 minutes to complete it. The participants were also asked to provide information regarding their age, gender, CGPA points, their perceptions of preparedness for the teaching profession, and for student assessment (refer to Chapter 3 for participant details).

5.1.2.3 Data Analysis

The study made use of a range of data analysis techniques, specifically psychometric analysis methods. The initial psychometric analysis, which was undertaken to provide a brief and explorative piece of information about the modified CAK measure, was 1PL model. After getting an initial sense of the data, CTT methodology was used to examine the relationship between item difficulty levels and discrimination indices. This technique also provided an examination of problematic items and a psychometric analysis of options. With the problematic items detected and eliminated, Rasch model analysis was carried out, which allowed a closer and more comprehensive review of the measure and its items. 2PL model for item analysis was also performed to inspect item discrimination properties. These procedures aimed at obtaining information regarding both measure properties and person abilities. As a last step of psychometric investigation, a Rasch PCA was carried out to seek any possible components or dimensions present in the measure.

Last analytical technique employed was correlations in order to explore the relationships among variables.

5.1.3 Overview of the Results

This study aimed to (a) investigate the psychometric properties of the modified CAK, (b) explore pre-service EFL teachers' LAL knowledge base and (c) find out about the factors contributing to the development of pre-service EFL teachers' LAL.

In relation to RQ1, the exploration of the psychometric properties of the modified CAK started with 1PL model in R, as the model provides a practical and comprehensive overview of the functioning of each item on an assessment instrument using item-person parameters and the their relationships. This was followed by a traditional item analysis based on CTT with the purpose of supporting the previous 1PL analysis and to have a better understanding of the functioning of any existing problematic items, receiving feedback from options analyses. The first two rounds of item and test analyses informed on three problematic items, which were then removed from the measure to go on with the rest of the analysis. With three problematic items removed from the measure, a second round of Rasch model analysis was performed, this time using the Winsteps® computer program (version: Ministeps) as it provides comprehensive and detailed Rasch model specifications supported with practical visuals. This round of Rasch analysis had a greater focus on logit values, item-person relationships and item-person map (Wright map) to better understand the positioning of items and persons relative to each other. The final type of analysis conducted with respect to RQ1 was a 2PL (IRT) model. This model was used to incorporate an IRT perspective into the item analysis, which involved the parameter of item discrimination in addition to the parameters of person ability and item difficulty. The results from this analysis produced similar outcomes to those from Rasch and CTT analyses. Lastly, Rasch PCA was performed to test dimenstionality of the instrument and its items.

In relation to RQ2, a closer look at the participant performances on the measure was taken, with particular interest in which items and which standards were found to be the most and the least difficult by the participants so as to have an idea about their strengths and weaknesses in terms of language assessment literacy. The results showed that the participants exhibited a moderate ability in language assessment literacy, and choosing the appropriate assessment methods was the easiest construct, while administering, scoring and interpreting assessment results was the most difficult. It was also noted that the greatest struggle of the participants was with two items which required them to employ some mathematical reasoning.

Finally, with respect to RQ 3, participant demographic information was explored with regard to its relationship with participant success in language assessment literacy. Of all the variables, CGPA, i.e., the general achievement in school subjects, was found to be the only variable to have a significant and positive correlation with total score from the measure; however, this correlation was not a large one, failing to explain much of the variance in the data.

5.2 Implications

This section provides the implications arising from the present study. The implications are divided into two categories: (a) psychometric implications and (b) pedagogical implications.

5.2.1 Psychometric Implications

As more emphasis is put on the role of assessment literacy in teaching, there is a continued search for using a valid and reliable assessment instrument to assess teachers for their assessment literacy. This study employed a modified and adapted version of the CAK instrument adapted by TAO (2014) based on Mertler and Campbell’s (2005) ALI instrument and on the Standards for Teacher Competence for Educational Assessment of Students (AFT, NCME & NEA, 1990). The measure was administered with the participation of a sample of 4th grade pre-service EFL teachers in Turkey. The data analysis (item and measure analyses) was performed using a combination of different perspectives to item and test analysis. Two model-fitting IRT approaches (1PL and 2PL), CTT approach and a Rasch model approach were used in combination; therefore, it could be argued that the present study adopted a comprehensive approach to item and test analyses in finding out about the psychometric properties of the measure.

Based on the psychometric analyses, three items were removed from the measure as they exhibited poor psychometric properties. This could be taken as a caution for future researchers who may wish to use the modified CAK instrument in their research. They may need to take this caution into consideration when recreating those items. On the other hand, the rest of the items on the instrument, and the overall instrument, could be reliably replicated in future research aiming to explore EFL teachers' language assessment literacy and their strengths and needs for training. However, it is recommended that the future reproductions of the instrument should employ its contextualised and tailored (based on needs) versions of the instrument as assessment literacy can be considered to be context-bound (Inbar-Lourie, 2008). The reproduction and recreation of the items should take into account the specifications outlined by the Standards. The reproduction of the instrument could also be done and used by teacher trainers and teacher training policy makers for both achievement and diagnostic purposes to inform instruction, policy and decision to arrive at such classroom-level and program-level decisions through a valid and reliable instrument that have operationalised internationally-recognised standards for assessment literacy into test items.

5.2.2 Pedagogical Implications

The literature both in language education and the broader educational sciences has continually shown that assessment is an integral part of teaching (Lukin, Bandalos, Eckhout, & Mickelson, 2004 Malone, 2013; Popham, 2009; Stiggins, 1999; White, 2009). However, there have been repeated reports of teachers' lacking necessary knowledge of and skills in assessment literacy (Fulcher, 2012; Inbar-Lourie, 2008, 2013; Lam & 2015, Taylor, 2013; Tsagari & Vogt, 2017; Volante & Fazio, 2007). It has also been argued that if properly conducted, classroom assessment not only informs but also enhances instruction (Black & William, 1998a, 1998b).

The participants in the present study were all 4th grade pre-service EFL teachers. They had already completed most of their course load and were at the point of graduating. In theory at least, they were supposed to be the group most representative of what and how pre-service EFL teachers learn about in the teacher training program. Therefore, an evaluation of their performance on the modified CAK instrument could suggest how their teacher training program prepared these students for assessment-related practices based on the *Standards*.

The results revealed that the mean success on the instrument was average. The majority of the participants were able to correctly answer only half of the items present on the instrument. No participant was able to answer all of the items correctly. There were only three participants who got 20 items right. These results become more serious considering the fact that all of the participants had already taken the English Language Testing course, and it had been less than a semester since they had taken the course at the time of data collection. Probably resulting from their own assessment experience as students and test-takers (given the testing culture in Turkey), the participants did not have serious problems with the items regarding choosing appropriate assessment methods. However, they had serious struggles in measurement-related concepts (such as percentile, raw score, and standard deviation). They also had serious problems in quality management of assessment procedures, and some fundamental principles of assessment such as validity, reliability, and types of assessment purposes). Therefore it could be argued that teacher training programs should put more emphasis on assessment in general, possibly offering more courses on assessment and testing, and should revise the curricula and course programs related to assessment in a way to increase the importance of such content areas as rationale behind the fundamental principles of assessment, making sense of and interpreting some of the basic mathematical and statistical procedures behind assessment, and more recent and innovative concepts in assessment.

5.3 Limitations and Future Directions

The present study aimed to investigate the psychometric properties of an instrument designed to assess EFL teachers' assessment literacy, which entailed a conceptual limitation because the instrument was capable of assessing only the knowledge base of assessment literacy. Even though the knowledge base of assessment literacy is a starting point of broader assessment literacy and has a close relationship with the ability to apply the knowledge of assessment into practice (Xu & Brown, 2017), assessment literacy is much more than its knowledge base. Therefore, a complete evaluation of language assessment literacy and the development of language assessment literacy would require going beyond the knowledge base and observe EFL teachers engaging with a variety of assessment practice both inside and outside the classroom in addition to a long-term monitoring of their assessment decisions to understand their impact on both short-term and long-term on student learning and achievement. Therefore, any academic inquiry in the future into language assessment literacy of EFL teachers can be recommended to take this gap into consideration.

Another limitation of the study relates to its methodology. As per the main research objective of this study, a quantitative research design was used by the study. However, the exploration of the second research objective definitely requires the use of a mixed research design to assess EFL teachers' language assessment literacy. Another methodological limitation concerned the sample size and diversity. Future studies are recommended increase both the sample size and diversity in a way to involve a more variety of settings in order to make generalizations about language assessment literacy of either pre-service or in-service EFL teachers.

Finally, in light of the considerations and arguments discussed in Chapter 4, the following suggestions regarding the future directions of LAL research in Turkey could be put forward. From a psychometric point of view, even though modified and adapted versions of the CAK instrument could be used with high validity and reliability confidence for the assessment and evaluation of EFL teachers' LAL at the knowledge base and with particular focus on the generic considerations in educational assessment and measurement, there is a need for the development of an instrument that could also provide accurate information on EFL teachers' language-specific AL with particular emphasis on the assessment of language skills and areas in addition to their knowledge of the basic concepts of assessment and measurement. Such an instrument must be reasoned, theory- and evidence-based, and psychometrically robust, evidenced through an in-depth qualitative and quantitative review of its validity and reliability. Extensive and comprehensive future research is also needed to better understand the causes of the seemingly chronic problem of poor LAL experienced by EFL teachers in Turkey as well as possible solutions to this problem through a systematic and thorough review of educational policies, English Language Teacher Education programs materials and curricula, and observations of the instruction of language testing and assessment courses taking into consideration both the teaching of theoretical knowledge and translation of that knowledge into practice by pre-service and in-service EFL teachers.

5.4 Conclusion

Similar to assessment literacy, language assessment literacy is gaining academic attention both around the world and in Turkey, with the use of assessment information continuing to change and evolve inside and outside the classroom. Not only teachers but also teacher training programs are at the forefront of this evolution and expected to adapt their policies accordingly. The present study investigated the psychometric properties of a modified version of CAK instrument, designed to be a measure of EFL teacher language assessment literacy, with the participation of a sample of 74 4th grade pre-service EFL teachers. The psychometric investigation involved item and test analyses (1PL and 2PL IRT models, Rasch analysis, and traditional CTT methodology). The investigation also included a Rasch PCA to test dimensionality within the measure to examine the component structure of the modified CAK to identify any present separate domains of the latent variable within the sample. Lastly, the relationships between background characteristics and the performance on the instrument were examined.

The results indicated that there were three items that were likely to produce faulty results and cause problems. They were eliminated, and the remaining items on the instrument implied presence of validity and reliability in the modified CAK, and indicated that the instrument could be replicated or reproduced with considerable confidence. The psychometric investigation also indicated a possibility of two separate components within the instrument. An examination of the participant performance on the instrument suggested that the sample had an average level of success overall, suggesting that they lacked adequate knowledge of assessment. Correlation analysis found that CGPA was the only statistically significantly correlation background characteristic with total score from the instrument. This study aimed to contribute to the body of literature related to EFL teachers' language assessment literacy. All possible stakeholders including teacher trainers, teacher training programs, pre-service and in-service teachers, and researchers should approach the exploratory results of the current study with caution, but can prefer to use this measure of language assessment literacy to inform the progression of pre-service teacher language assessment knowledge, to inform policy and decisions and to monitor self-progress.

REFERENCES

Alkharusi, H. (2011). Teachers' classroom assessment skills: Influence of gender, subject area, grade level, teaching experience and in-service assessment training. *Journal of Turkish Science Education, 8*(2), 39-48.

Alkharusi, H., Kazem, A. M., & Al-Musawai, A. (2011). Knowledge, skills, and attitudes of preservice and in-service teachers in educational measurement. *Asia-Pacific Journal of Teacher Education, 39* (2), 113-123. https://doi.org/10.1080/1359866X.2011.560649

Alkharusi, H., Aldhafri, S., Alnabhani, H., & Alkalbani, M. (2012). Educational assessment attitudes, competence, knowledge, and practices: An exploratory study of Muscat teachers in the Sultanate of Oman. *Journal of Education and Learning, 1*(2), 217-232.

Al-Malki, M. A. & Weir, K. (2014). A comparative analysis between the assessment criteria used to assess graduating teachers at Rustaq College (Oman) and Grifth University (Australia) during the teaching practicum. *Australian Journal of Teacher Education, 39*(12), 28-42. https://doi.org/10.14221/ajte.2014v39n12.3

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1), I-7.

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

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: Author.

Athanases, S. Z., Bennett, L. H., & Wahleithner, J. M. (2013). Fostering data literacy through preservice teacher inquiry in English language arts. *The Teacher Educator*, *48*(1), 8-28.

Bailey, K. M. (2001). Teacher preparation and development. *TESOL Quarterly, 35* (4), 609- 616. http://dx.doi.org/10.2307/3588439

Baker, F. B. (1985). *The basics of item response theory.* Portsmouth, NH: Heinemann

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Bastian, K. C., Henry, G. T., Pan, Y., & Lys, D. (2016). Teacher candidate performance assessments: Local scoring and implications for teacher preparation program improvement. *Teaching and Teacher Education, 59*, 1-12. http://dx.doi.org/10.1016/j.tate.2016.05.008

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

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*(1), 1-42.

Bachman, L. F. (2014). Ongoing challenges in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (1st ed., Vol. 3), (pp. 1586-1603). Oxford: John Wiley and Sons.

Betebenner, D. (2009). Norm and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42-51.

Beziat, T. L. R., & Coleman, B. K. (2015). Classroom assessment literacy: Evaluating preservice teachers. *The Researcher*, *27*(1), 25-30.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17-66). New York: Springer.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy, & Practice, 5*(1), 7-73.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *The Phi Delta Kappan, 80*(2), 139-148.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences.* New York, NY: Routledge.

Boone, W. J., Staver, J.R., & Yale, M. S. (2013). *Rasch analysis in the human sciences.* Springer Science & Business Media.

Boud, D. (2006). Forward. In C. Bryan & K. Clegg (Eds.), *Innovative assessment in higher education* (pp. xvii-xix). New York: Routledge.

Boyles P. (2005). Assessment literacy. In Rosenbusch M. (Ed.), *National Assessment Summit Papers,* 11–15. Ames, IA: Iowa State University.

Briggs, S. R., & Cheek, J. M. (1988). On the nature of self-monitoring: Problems with assessment, problems with validity. *Journal of Personality and Social Psychology, 54*(4), 663.

Brookhart, S. M. (2001). The "Standards" and classroom assessment aresearch. *Educational Measurement: Issues and Practice, 18*(1), 23-27.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30*(1), 3-12.

Brown, H. D. (2004). *Language Assessment: Principles and Classroom Practices*. New York: Pearson/Longman, 2004.

Chan, Y. C. (2008). Elementary school EFL teachers' beliefs and practices of multiple assessments. *Reflections on English Language Teaching, 7*(1), 37–62.

Calfee, R. C., & Masuda, W. V. (1997). Classroom assessment as inquiry. In G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, adjustment, and achievement.* New York: Academic Press.

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford: Oxford University Press.

Carroll, B. J. (1980) Specifications for an English language testing service. In Alderson, J. C. and Hughes, A. (eds) *Issues in Language Testing*. ELT Documents 111. London: British Council, 66–110.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254- 27 2.

Chapman, M. L. (2008). *Assessment literacy and efficacy: Making valid educational decisions*. Unpublished doctoral thesis. University of Massachusetts Amherst.

Cheng, L. (2001). An investigation of ESL/EFL teachers' classroom assessment practices. *Language Testing Update, 29,* 53-64.

Cheng, L., Rogers, T., & Hu, H. (2004). esl/efl instructors' classroom assessment practices: Purposes, methods, and procedures. *Language Testing, 21*(3) 360-389. https:// doi.org/10.1191/0265532204lt288oa.

Clark-Gareca, B. (2016). Classroom assessment and English Language Learners: Teachers' accommodations implementation on routine math and science tests. *Teaching and Teacher Education, 54*, 139-148. http://dx.doi.org/10.1016/j.tate.2015.11.003

Coombe, C. (2018). *An A to Z of Second Language Assessment: How Language Teachers Understand Assessment Concepts*. London, UK: British Council.

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

Conor, U. & Mbaye, A. (2002). Discourse approaches to writing assessment. *Annual Review of Applied Linguistics, 22*, 263-278.

Creswell, J. W. (2014). *Research design: qualitative, quantitative, and mixed methods approaches.* Thousand Oaks: SAGE Publications.

Cronbach, L. J. 1988. Five perspectives on validity argument. In H. Wainer & H. I. Braun (eds.). *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum Associates, 3-17.

Cronbach, L. J. and Meehl, P. E. (1955) Construct validity in psychological tests. *Psychological Bulletin* 52, 281–302.

Crystal, D. (2003). *English as a global language.* (2nd ed.). Cambridge: Cambridge University Press.

Daniel, L. G., & King, D. A. (1998). Knowledge and use of testing and measurement literacy of elementary and secondary teachers. *The Journal of Educational Research, 9*1(6), 331-344.

Darling-Hammond, L., & Adamson, F. (2013). *Developing assessments of deeper learning: The costs and benefits of using tests that help students learn.* Stanford, California: Stanford Center for Opportunity Policy in Education.

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

Dayal, H. C., & Lingam, G. I. (2015). Fijian Teachers' Conceptions of Assessment. *Australian Journal of Teacher Education, 40*(8), 43-58. http://dx.doi.org/10.14221/ajte.2015v40n8.3

DeAyala, R. J. (2013). *The theory and practice of item response theory.* New York, NY: Guilford Publications.

DeLuca, C., & Bellara, A. (2013). The current state of assessment education: Aligning policy, standards, and teacher education curriculum. *Journal of Teacher Education, 64*(4), 356-372.

DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice, 17*(4), 419-438.

Díaz, C., Alarcón, P., & Ortiz, M. (2012). El profesor de inglés: sus creencias sobre la evaluación de la lengua inglesa en los niveles primario, secundario y terciario [The English teacher: His beliefs about English language assessment at primary, secondary and tertiary levels]. Íkala, Revista de Lenguaje y Cultura, 17(1), 15-26.

Dimitrov, D. M. (2013) *Quantitative research in education: Intermediate & advanced methods.* Oceanside, NY: Whittier Publications.

Dinther, M. V., Dochy, F., & Segers, M. (2015). The contribution of assessment experiences to student teachers' self-efficacy in competence-based education. *Teaching and Teacher Education, 49*, 45-55.

Earl, L. M. (2013). *Assessment as learning: Using classroom assessment to maximize student learning* (2nd ed.). Thousand Oaks, California: Corwin.

Falchikov, N., & Boud, D. (2008). The role of assessment in preparing for lifelong learning: Problems and challenges. In A. Havnes & L. McDowell (Eds.), *Balancing dilemmas in assessment and learning in contemporary education* (pp. 87-99). New York: Routledge.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-381.

Field, A. P. (2009). *Discovering statistics using IBM SPSS Statistics.* (3rd. Ed.). London: SAGE

Fishbein, M., & Ajzen, I. (2010). P*redicting and changing behavior: The reasoned action Approach*. New York: Psychology Press (Taylor & Francis).

Freeman, D. (2001). Second language teacher education. In R. Carter & D. Nunan (Eds.), *The Cambridge guide to teaching English to speakers of other languages* (pp. 72-79). Cambridge: Cambridge University Press.

Fulcher, G. (1999) Assessment in English for academic purposes: putting content validity in its place. *Applied Linguistics 20*(2), 221–236.

Fulcher, G., & Davidson, F. (2012). *Language testing and assessment: an advanced resource book*. Abingdon: Routledge.

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

Gareis, C. R., & Grant, L. W. (2015). Assessment literacy for teacher candidates: A focused approach. *Teacher Educators' Journal,* 4-21.

Gazi University Department of English Language Teaching. (n.d.). *Gazi Üniversitesi Gazi Eğitim Fakültesi Yabancı Diller Eğitimi Bölümü İngiliz Dili Eğitimi Anabilim Dalı lisans programı ders içerikleri*. Retrieved November 15, 2019 from http://gazi.edu.tr/posts/download?id=211331

Gotch, C. M. (2012). *An investigation of teacher educational measurement literacy.* Unpublished Doctoral Dissertation, Washington State University, Washington.

Gotch, C. M., & French, B. F. (2013). Elementary teachers' knowledge and self-Efficacy for measurement concepts. *The Teacher Educator 48*(1), 46-57.

Green, A. (2013). *Exploring language assessment and testing: Language in action*. New York: Routledge.

Griffin, P., Care, E., & McGaw, B. (2012). The changing role of education and schools. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 1-15). New York: Springer.

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

Gummer, E. S., & Mandinach, E. B. (2015). Building a conceptual framework for data literacy. *Teachers College Record, 117*(4), 1-22.

Harding, L., & Kremmel, B. (2016). Teacher assessment literacy and professional development. In D. Tsagari & J. Banerjee (Eds.), *Handbooks of applied linguistics [HAL]: Handbook of second language assessment* (pp. 413-427). Berlin, Germany/Boston, MA: De Gruyter Mouton.

Hasselgreen, A., Carlsen, C., & Helness, H. (2004). *European survey of language testing and assessment needs. Part 1: General findings.* Gothenburg, Sweden: European Association for Language Testing and Assessment.

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. S. Sifakis (Eds). *English Language Education Policies and Practices in the Mediterranean Countries and Beyond* (pp. 227-257). Frankfurt: Peter Lang.

Heaton, J. (1990). *Writing English language tests*. New York: Longman Inc.

Heritage, M. (2013). *Formative assessment in practice: A process of inquiry and action.* Cambridge, Massachusetts: Harvard Education Press.

Hughes, A. (1989). *Testing for language teachers* (2nd ed.). New York: Cambridge University Press.

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

Inbar-Lourie, O. (2013). Language assessment literacy. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics (pp.1-9).* Blackwell Publishing Ltd.

Inbar-Lourie, O. (2017). Language assessment literacy. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (pp. 1–14). New York, NY: Springer.

Jimerson, J. B., & Wayman, J. C. (2015). Professional learning for using data: Examining teacher needs and supports. *Teachers College Record, 117*(4), 1-36.

Joughin, G. (2009). Assessment, learning and judgment in higher education: A critical review. In G. Joughin (Ed.), *Assessment, learning and judgment in higher education* (pp. 1-15). New York: Springer.

Kahl, S. R., Hofman, P., & Bryant, S. (2013). *Assessment literacy standards and performance measures for teacher candidates and practicing teachers* . Dover, NH: Measured Progress.

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin.* 12, 527- 535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement 38*(4), 319-342.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport, CT: Praeger.

King, J. D. (2010). *Criterion-referenced assessment literacy of educators*. Unpublished doctoral thesis. The University of Southern Mississippi.

Kırkgöz, Y. (2009). Globalization and English language policy in Turkey. *Educational Policy, 23* (5), 663-684.

Kline, R. B. (2015). *Principles and practice of structural equation modeling.* Guilford publications.

Kunnan, A. J. (2000) Fairness and justice for all. In Kunnan, A. J. (ed.) *Fairness and Validation in Language Assessment*. Studies in Language Testing. Cambridge: Cambridge University Press, 1–14.

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

Lamprianou, I., & Athanasou, J. A. (2009). *A teacher's guide to educational assessment* (Revised edition). Rotterdam: Sense.

Leighton, J. P. (2011). Editorial. *Educational Measurement: Issues and Practice 30*(1), 1-2.

Linacre, J. M. (1998). Structure in Rasch residuals: Why principal components analysis? *Rasch Measurement Transactions, 12*(2), 636.

Linacre J. M., Wright B. D. (2000). *A user's guide to Winsteps Rasch-model computer programs*. (Unpublished technical report). Chicago, IL: MESA Press.

Linacre, J. M. (2012). *Winsteps Rasch Tutorial 2* (Unpublished technical report). Retrieved from: http://www.winsteps.com/tutorials.html.

López, A., & Bernal, R. (2009). Language testing in Colombia: A call for more teacher education and teacher training in language assessment. *Profile: Issues in Teachers' Professional Development, 11*(2), 55-70.

Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice, 23*(2), 26-32.

Maclellan, E. (2004). Initial knowledge states about assessment: novice teachers' conceptualisations. *Teaching and Teacher Education, 20*, 523–535. https://doi.org/10.1016/j.tate.2004.04.008

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Masters, G. N. (2013). *Reforming educational assessment: Imperatives, principles and challenges.* Australian Education Review 57. Melbourne: Australian Council for Educational Research.

McMillan, J. H., & Nash, S. (2000). Teacher classroom assessment and grading practices decision making. Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans.

McMillan, J. H. (2001). *Essential assessment concepts for teachers and administrators.* Thousand Oaks, CA: Corwin Publishing Company.

McMillan, J. H. (2014). *Classroom assessment: Principles and practice for effective standards-based instruction* (6th ed.). Boston: Pearson Education.

McNamara, T. (2006). Validity in language testing: the challenge of Sam Messick's legacy. *Language Assessment Quarterly 3* (1), 31–51.

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

Mellati, M., Khademi, M., & Shirzad, A. (2015). The Relationships among sources of teacher pedagogical beliefs, teaching experiences, and student outcomes. *International Journal of Applied Linguistics & English Literature, 4*(2), 177-184. http://dx.doi.org/10.7575/aiac.ijalel.v.4n.2p.177

Mertler, C. A. (1999). Assessing student performance: A descriptive study of the classroom assessment practices of Ohio teachers. *Education*, 120, 285-296.

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 their perceptions of the impact of classroom assessment professional development. *Improving Schools, 12*(2), 101-113.

METU. (n.d). *Academic catalog.* Retrieved November 15, 2019, from https://catalog.metu.edu.tr/course.php?prog=450&course_code=4500413)

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). Old Tappan, NJ: MacMillan.

Mertler, C.A. & Campbell, C.S. (2005). Measuring teachers' knowledge and application of classroom assessment concepts: Development of the Assessment Literacy Inventory. Proceeding from *American Educational Research Association*, Montreal, Quebec, Canada.

Molloy, E., & Boud, D. (2014). Feedback models for learning, teaching and performance. In J.M. Spector, M.D. Merrill, J. Elen, & M.J. Bishop, (Eds.), *Handbook of Research on Educational Communications and Technology* (pp. 413-424). New York: Springer.

Moss, C. M., & Brookhart, S. M. (2012). *Learning targets: Helping students aim for understanding in today's lesson.* Al xandria, Virginia: Association for Supervision and Curriculum Development.

Mousavi, S. A. (2002). *An encylopedic dictionary of language testing*. (3rd edition). Taiwan: Tung Hua Book Company.

Mujis, D. (2004). *Doing quantitative research in education with SPSS.* SAGE Publications Ltd.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. J.W. Pellegrino, N. Chudowsky & R. Glaser, (Eds.). Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, D.C.: National Academy Press.

Nicol, D. (2013). Resituating feedback from the reactive to the proactive. In D. Boud & E. Molloy (Eds.), *Feedback in higher and professional education: Understanding it and doing it well* (pp. 34-49). New York: Routledge.

Nunan, D. (2001). English as a global language. *TESOL Quarterly, 35* (4), 605-606. http://dx.doi.org/10.2307/3588436

Ölmezer-Öztürk, E. & Aydın, B. (2018). *Developing and Validating Language Assessment Knowledge Scale (LAKS) and Exploring the Assessment Knowledge of EFL Teachers.* Unpublished PhD Dissertation, Anadolu University, Turkey.

Ölmezer-Öztürk, E. & Aydin, B. (2018). Investigating language assessment knowledge of EFL teachers. *H.U. Journal of Education.*

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

Ozturk, U., & Atay, D. (2010). Challenges of being a non-native English teacher. *Educational Research.* 1, 135-139.

Pill, J., & 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 assessment of students. *Mid-Western Educational Researcher, 6* (1), 21-27.

Popham, W. J. (1995). *Classroom assessment: What teachers need to know.* Boston: Allyn-Bacon.

Popham, W. J. (2004). Why assessment illiteracy is professional suicide. *Educational Leadership, 62*(1), 82.

Popham, W. J. (2006). All About Accountability: A Dose of Assessment Literacy. *Improving Professional Practice, 63*(6), 84–85.

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

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator, 46*(4), 265-273.

Popham, W. J. (2014). *Classroom assessment: What teachers need to know* (7th ed.). Boston: Pearson Education.

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

Purpura, J. E. (2016). Second and foreign language assessment. *Modern Language Journal, 100* (Supplement 2016), 290-208. (Centennial Issue).

Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics research. *Language Learning, 65*(S1), 37–75.

Quilter, S. M., & Gallini, J. K. (2000). Teachers’ assessment literacy and attitudes. *The Teacher Educator, 36*(2), 115 - 131.

Rea-Dickins, P. (2007). Classroom-based assessment: Possibilities and pitfalls. In J. Cummins, & C. Davison (Eds.), *International Handbook of English Language Teaching* (pp. 505-520). New York: Springer.

Reise, S. P., & Revicki, D. A. (2015). *Handbook of item response theory modeling: applications to typical performance assessment.* New York: Routledge, Taylor & Francis Group.

Richards, J. (2008). Second language teacher education today. *RELC Journal, 39* (2), 158- 177. https://doi.org/10.1177/0033688208092182

Richards, J., & Nunan, D. (1990). *Second language teacher education*. Cambridge: Cambridge University Press.

Rizopoulos, D. (2006) ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1–25. URL http://www.jstatsoft.org/v17/ i05/

Russell, M. K., & Airasian, P. W. (2012). *Classroom assessment: Concepts and applications* (7th ed.). New York: McGraw-Hill.

Ryaner, K. A. (2018). *An investigation of pre-service teacher assessment literacy and assessment confidence: Measure development and EDTPA performance*. Unpublished doctoral thesis. Kent State University.

Şahin, M. (2019). *An analysis of English language testing and evaluation course in English language teacher education programs in Turley: developing language assessment literacy of pre-service EFL teachers.* Unpublished doctoral thesis. METU.

Sanders, J. R., & Vogel, S. R. (1993). The development of standards for teacher competence in educational assessment of students. In S. L. Wise (Ed.), *Teacher training in measurement and assessment skills,* Lincoln, NB: Burros Institute of Mental Measurements.

Schafer, W. D., & Lizzitz, R. W. (1987). Measurement training for school personnel: Recommendations and reality. *Journal of Teacher Education, 38*(3), 57-63.

Scott, S., Webber, C. F., Aitken, N., & Lupart, J. (2011). Developing teachers' knowledge, beliefs, and expertise: Findings from the Alberta Student Assessment Study. *The Educational Forum, 75*(2), 96–113. https://doi.org/10.1080/00131725.2011.552594

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education* (19), 405- 450.

Shepard, L. A. (2008). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of assessment: Shaping, teaching and learning* (pp. 279-303). New York: Lawrence Erlbaum Associates.

Shohamy, E. (2001). T*he power of tests: A critical perspective on the uses of language tests.* Harlow, England: Pearson Education Limited.

Shute, V. J. (2008). Focus on formative feedback. Review of educational research, 78(1), 153-189.

Siegel, M. A., & Wissehr, C. (2011). Preparing for the plunge: Preservice teachers' Assessment Literacy. *Journal of Science Teacher Education, 22*(4), 371-391.

Stabler-Havener, M. L. (2018). Defining, conceptualizing, problematizing, and assessing language teacher assessment literacy. *Working Papers in Applied Linguistics & TESOL, 18*(1), 1-22

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. (2006). Assessment for Learning: A Key to Student Motivation and Learning. *Phi Delta Kappa Edge, 2*(2), 1–19.

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 & Francis.

Stiggins, R. J., & Conklin, N. F. (1992). *In teachers' hands: Investigating the practices of classroom assessment.* Albany, NY: State University of New York Press, Albany.

Supovitz, J. (2010). Knowledge-based organizational learning for instructional improvement. In *Second international handbook of educational change*, 701-723. Netherlands: Springer.

Susuwele-Banda, W.J. (2005). *Classroom Assessment in Malawi: Teachers' Perceptions and Practices in Mathematics.* Unpublished Doctoral Dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Szabó G. (2012). *Applying item response theory in language test item bank building.* Frankfurt a.M.: Peter Lang GmbH, Internationaler Verlag der Wissenschaften.

Taber, K. S. (2018). The use of Cronbach's Alpha when developing and reporting research instruments in science education. *Res Sci Educ* 48, 1273–1296 (2018). https://doi.org/10.1007/s11165-016-9602-2

Tang, K. C. C. (1994). Assessment and student learning: Effects of modes of assessment on students‟ preparation strategies. In G. Gibbs (Ed.), *Improving student learning: Theory and practice* (pp. 151-170). Oxford: The Oxford Centre for Staff Development.

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

Taylor, C. S., & Nolen, S. B. (2008). *Classroom assessment: Supporting teaching and learning in real classrooms* (2nd ed.). Upper Saddle River, New Jersey: Pearson Education.

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.

Thorndike, R. L., & Hagen E. (1969). *Measurement and Evaluation in Psychology and Education (3rd ed.)*. New York: John Wiley and Sons.

Vogt, K., & Tsagari, D. (2014) Assessment Literacy of Foreign Language Teachers: Findings of a European Study, *Language Assessment Quarterly, (11)*4, 374-402, DOI: 10.1080/15434303.2014.960046

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

Volante, L., & Fazio, X. (2007). Exploring teacher candidates' assessment literacy: Implications for teacher education reform and professional development. *Canadian Journal of Education, 30*(3), 749-770. https://doi. org/10.2307/20466661.

Walsh, E., & Isenberg, E. (2015). How does value added compare to student growth percentiles? *Statistics and Public Policy, 2*(1), 1-13.

Wang, T. H., Wang, K. H., & Huang, S. C. (2008). Designing a web-based assessment environment for improving pre-service teacher assessment literacy. *Computers & Education, 51*(1), 448-462.

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

Widdowson, H. G. (1997). Approaches to second language education. In G. R. Tucker & D. Corson (Eds.), *Encyclopedia of language and education: Volume 4* (pp. 123-128). Dordrecht: Kluwer Academic Publishers

Wiggins, G. (1998). *Educative Assessment: Designing Assessments to Inform and Improve Student Performance*. San Francisco, California. Jossey-Bass.

Wiliam, D., & Thompson, M. (2008). Integrating assessment with learning: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment.* New York: Lawrence Erlbaum Associates.

Wright, B. D. (1996). Local dependency, correlations and principal components. *Rasch Measurement Transactions*, *10*(3), 509-511.

Wright, T. (2010). Second language teacher education: Review of recent research on practice. *Language Teaching, 43* (3), 259-296.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement.* Mesa Press. Chicago, IL: Mesa Press.

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach.* Melbourne: Educational Measurement Solutions.

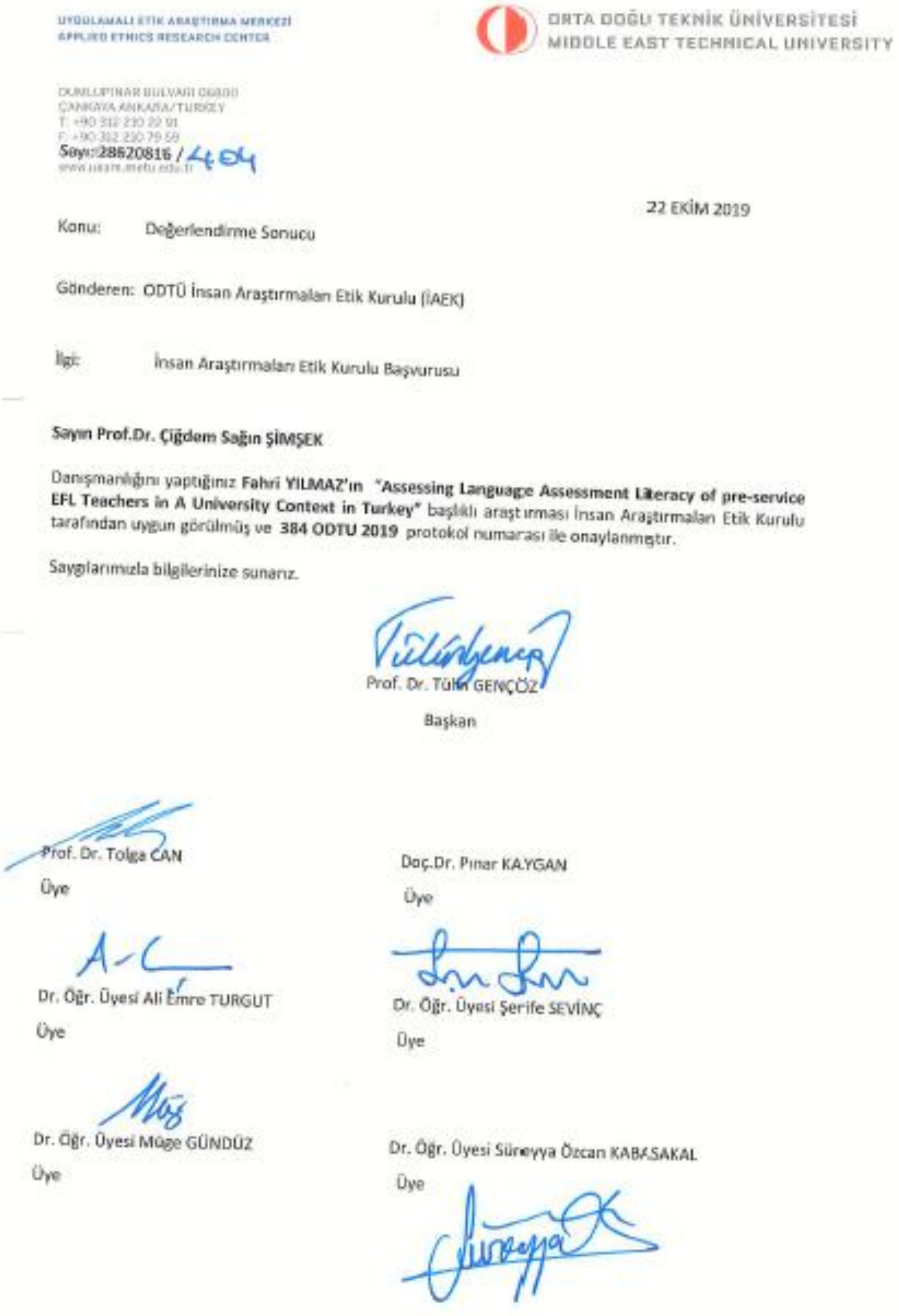
Xu, Y., & Brown, G.T.L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, 58, 149-162.

Xu, Y., & Brown, G. T. L. (2017). University English teacher assessment literacy: A survey-test report from China. *Papers in Language Testing and Assessment, 6* (1), 133-158.

Yan, Z. & Cheng, E. C. K. (2015). Primary teachers' attitudes, intentions and practices regarding formative assessment. *Teaching and Teacher Education, 45*, 128-136. http://dx.doi.org/10.1016/j.tate.2014.10.002

APPENDICES

A. HUMAN SUBJECTS ETHICS COMMITTEE APPROVAL



UYGULAMALI ETİK ARAŞTIRMA MERKEZİ
APPLIED ETHICS RESEARCH CENTER

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

DUMLUPINAR BULVARI 06800
ÇANKAYA ANKARA/TURKEY
T: +90 312 210 22 91
F: +90 312 210 79 59
Sayı: 28620816 / 404
www.ueam.metu.edu.tr

22 EKİM 2019

Konu: Değerlendirme Sonucu

Gönderen: ODTÜ İnsan Araştırmaları Etik Kurulu (İAEK)

İlgi: İnsan Araştırmaları Etik Kurulu Başvurusu

Sayın Prof.Dr. Çiğdem Sağın ŞİMŞEK

Danışmanlığını yaptığınız **Fahri YILMAZ'ın "Assessing Language Assessment Literacy of pre-service EFL Teachers in A University Context in Turkey"** başlıklı araşt ırması İnsan Araştırmaları Etik Kurulu tarafından uygun görülmüş ve **384 ODTU 2019** protokol numarası ile onaylanmıştır.

Saygılarımızla bilgilerinize sunarız.

Prof. Dr. Tülin GENÇÖZ
Başkan

Prof. Dr. Tolga CAN
Üye

Doç.Dr. Pınar KAYGAN
Üye

Dr. Öğr. Üyesi Ali Emre TURGUT
Üye

Dr. Öğr. Üyesi Şerife SEVİNÇ
Üye

Dr. Öğr. Üyesi Müge GÜNDÜZ
Üye

Dr. Öğr. Üyesi Süreyya Özcan KABASAKAL
Üye

B. DEBRIEFING FORM

This study aims to collect data in order to (a) explore the language assessment literacy of pre-service EFL teachers following a course on English language testing and to (b) find out about the validity of a language assessment knowledge scale.

Language Assessment Literacy, which originated from a broader concept of Assessment Literacy, refers to language teachers' familiarity with the basic concepts and principles of language testing and assessment and the application of this knowledge to classroom practices in general and specifically to issues related to language assessment. There is a general consensus among researchers that teachers with a thorough understanding of assessment literacy can make sophisticated and informed decisions and judgments about the validity and reliability of practices and policies related to assessment in certain contexts. The importance of assessment knowledge is reflected in several authors' writings who state that assessment is half of teaching.

However, assessment literacy, and assessment in general, are often ignored in the Turkish context, not unlike the situation around the world, both in the pre-service education and in-service training programs. Therefore, it is important to know pre-service EFL teachers' assessment knowledge and to know how best to assess language assessment knowledge. This research study intends to serve those purposes as well as providing an opportunity to investigate the syllabus of a language testing course as the assessment literacy data will be obtained from pre-service EFL teachers who just taken a course on language testing.

It is aimed that the data from this study will be obtained at the end of December 2019. These data will be utilized only for research purposes. For further information, about the study and its results, you can refer to the following contact information. I would like to thank you for participating in this study.
Fahri Yılmaz (Tel: 0506 739 2694; E-mail: fahri.yilmaz@metu.edu.tr)

C. INFORMED CONSENT FORM

The aim of this study, carried out by Fahri Yılmaz, an MA student at the METU ELT Department, is to collect data in order to (a) explore the language assessment literacy of pre-service EFL teachers and to (b) find out about the validity of a language assessment knowledge scale. Participation in the study must be on a voluntary basis. No personal identification information is required in the questionnaire. Your answers will be kept strictly confidential and evaluated only by the researcher; the obtained data will be used for scientific purposes.

The questionnaire does not contain any questions that may cause discomfort in the participants. However, during participation, for any reason, if you feel uncomfortable, you are free to quit at any time. In such a case, it will be sufficient to tell the person conducting the survey (i.e., data collector) that you haven not completed the questionnaire.

After all, the questionnaires are collected back by the data collector, your questions related to the study will be answered. I would like to thank you in advance for your participation in this study. For further information about the study, you can contact Fahri Yılmaz; E-mail: fahri.yilmaz@metu.edu.tr).

I am participating in this study totally on my own will and am aware that I can quit participating at any time I want/ I give my consent for the use of the information I provide for scientific purposes. (Please return this form to the data collector after you have filled it in and signed it).

Name Surname Date Signature

D. ORIGINAL CAK

DIRECTIONS *The following items are examining your knowledge in the educational assessment of students. Please read each scenario followed by each item carefully and answer each of the items by **circling** the response you think is the best one. **Even if you are not sure of your choice, circle the response you believe to be the best. Do not leave any items unanswered.*** **Scenario # 1** Mr. Chan Sambath, a first year English writing lecturer, is aware of the fact that his students will be taking a semester examination at the end of the course.

1. Mr. Chan Sambath wants to assess his students‟ critical thinking abilities at the end of the unit to determine if any reinstruction will be necessary prior to the exam. Which of the following methods would be the **most** appropriate choice?

A. multiple-choice items
B. matching items
C. gap-filling items
D. essay writing

2. In order to grade his students‟ writing accurately and consistently, Mr. Chan Sambath would be **well** advised to

A. identify criteria from the unit objectives and create a marking criteria.
B. develop a marking criteria after getting a feel for what students can do.
C. consider student performance on similar types of tests.
D. consult with experienced colleagues about a marking criteria that has been used in the past.

3. Mr. Chan Sambath wants to evaluate his students‟ understanding of specific aspects of their responses. Which of the following would **best** facilitate him scoring of these responses?

A. an objective answer key
B. an holistic scoring
C. a checklist
D. an analytic scoring

4. At the end of each class period, Mr. Chan Sambath asks his students several questions to get an impression of their understanding. In this example, the **primary** purpose for conducting formative assessment is to

A. determine the final grades for students.
B. determine content for the final examination.
C. identify individual learning needs to plan classroom instruction.

D. evaluate curriculum appropriateness.
5. Which grading practice being considered by Mr. Chan Sambath would result in grades that would **most** reflect his students‟ learning achievement against the learning outcomes?

A. grades based on the students‟ performances on a range of assessments
B. grades based on the amount of time and effort the student spent on the assessments
C. grades based on how the student has performed in comparison to his/her classmates
D. grades based upon the personal expectations of Mr. Chan Sambath

6. Mr. Chan Sambath is planning to keep assessment records as a part of his assessment and reporting process. Which of the following is the **least** important assessment information to be recorded?

A. statistical data including marks, student welfare and biographical information.
B. anecdotal data comprising critical incidents or reflections of both Mr. Chan Sambath and his students.
C. all copies of his students‟ assessment work.
D. a representative sample of each student work.

7. In a routine conference with his students, Mr. Chan Sambath is asked to explain the **basis** for assigning his course grade. Mr. Chan Sambath should

A. explain that the grading system was imposed by the school administrators.
B. refer to the information that he presented to his students at the beginning of the course
on the assessment process.
C. re-explain the students the way in which the grade was determined and show them
samples of their work.
D. indicate that the grading system is imposed by the Ministry of Education.

8. Mr. Chan Sambath was worried that his students would not perform well on the semester examination. He did all of the following to help increase his students‟ scores. Which was **unethical**?

A. He instructed his students in strategies for taking tests.
B. He planned his instruction so that it focused on concepts and skills to be covered on the test.
C. He allowed his students to bring in their coursebooks/materials to refer to during the test.
D. He allowed students to practice with a small number of items from the actual test.

9. To ensure the validity and reliability of his classroom assessment procedure, it is advised that Mr. Chan Sambath should gather together with his colleagues to discuss all of the following **except**

A. marking criteria.
B. students‟ pieces of work.
C. teaching techniques.
D. assessment activities.

Scenario # 2 Ms. Chan Tevy is a year two English lecturer. She has just finished teaching a unit on the Industrial Revolution and wishes to measure her students‟ understanding of this particular unit using a multiple-choice test.

10. Based on her goal, which of the following assessment strategies would be the **most** appropriate choice?

A. She should use the test items included in the teacher‟s manual from the textbook she uses.
B. She should design test items which are consistent with the content and skill specified
in the course learning outcomes.
C. She should use available test items from internet that cover Industrial Revolution.
D. She should design test items which cover the factual information she taught.
11. In constructing her multiple-choice test items, Ms. Chan Tevy should follow all of the following guidelines **except**

A. ensure that the correct response is unequivocally the best.

B. ensure that the responses to a given item are in different literary forms.

C. ensure the stem and any response, taken together, read grammatically.

D. make all distracters plausible and attractive to the ignorant test-taker.

12. Ms. Chan Tevy decides to score the tests using a 100% correct scale. Generally speaking, what is the **proper** interpretation of a student score of 85 on this scale?

A. The student answered 85% of the items on the test correctly.
B. The student knows 85% of the content covered by this instructional unit.
C. The student scored higher than 85% of other students who took this test.
D. The student scored lower than 85% of other students who took this test.

13. Some of Ms. Chan Tevy‟s students do not score well on the multiple-choice test. She decides that the next time she teaches this unit, she will begin by administering a pretest to check for students‟ prerequisite knowledge. She will then adjust her instruction based on the pretest results. What **type** of information is Ms. Chan Tevy using?

A. norm-referenced information (describes each student‟s performance relative to the other students in a group such as percentile ranks)
B. criterion-referenced information (describes each student‟s performance in terms of status in specific learning outcomes)
C. both norm- and criterion-referenced information
D. neither norm- nor criterion-referenced information

14. The Industrial Revolution test is the only student work that Ms. Chan Tevy grades for the current grading period. Therefore, grades are assigned only on the basis of the test. Which of the following is **not** a criticism of this practice?

A. The test, and therefore the grades, reflect too narrow a curriculum focus.
B. These grades, since based on test alone, are probably biased against some minority students.
C. Tests administered under supervised conditions are more reliable than those assessments undertaken in less standardized conditions (e.g. homework)
D. Decisions like grades should be based on more than one piece of information.

15. Ms. Chan Tevy fully understands that her classroom assessment records serve all of the following purposes **except**

A. provide information regarding assessment methods development.
B. provide diagnostic information to show the strengths and weaknesses of student performance.
C. show the extent of student progress.
D. provide information to assist administrative decision makers.

16. During an individual conference, one student in Ms. Chan Tevy‟s class wants to know what it means that he scored in the 80th percentile in a multiple-choice test. Which of the following provides the **best** explanation of this student‟s score?

A. He got 80 % of the items on the test correct.
B. He is likely to earn a grade of “B” in his class.
C. He is demonstrating above grade level performance.
D. He scored the same or better than 80 % of his classmates.

17. Based on their grades from last semester, Ms. Chan Tevy believes that some of her low-scoring students are brighter than their test scores indicate. Based on this knowledge, she decides to add some points to their test scores, thus raising their grades. Which of Ms. Chan Tevy‟s action was **unethical**?

A. examining her student‟s previous academic performance
B. adjusting grades in her course
C. using previous grades to adjust current grades
D. adjusting some students‟ grades and not others‟

18. To enhance the quality of a new developed multiple-choice test, Ms. Chan Tevy should do all of the following **except**

A. pilot the test items with a small number of her past students to see how well each item performs.
B. make all necessary changes to the test items based on the information received during her pilot.
C. have all of her current students undertake the test twice and make a comparison of their scores.
D. panel the test items through consultation with her colleagues who have assessment experience.

Scenario # 3 Mr. Peo Virak is a senior English lecturer in the Indrak Tevy University. Experienced in issues of classroom assessment, Mr. Peo Virak is often asked to respond to the questions concerning best practices for evaluating student learning.

19. Ms. Meas Chakriya, an English lecturer, asks what type of assessment is best to determine how well her students are able to apply what they have learned in class to a situation encountered in their everyday lives. The type of assessment that would **best** answer her question is called

A. diagnostic assessment.
B. performance assessment.
C. formative assessment.
D. authentic assessment.

20. Ms. Keo Bopha is constructing essay questions for a test to measure her students‟ critical thinking skills. She consults with Mr. Peo Virak to see what concerns she would be aware of when constructing the questions. Which statement is **not** an appropriate recommendation when writing essay questions?

A. consider the relevance of the questions for a particular group of her students
B. avoid determining the amount of freedom of writing responses that will be accepted
C. indicate the time limits for the writing responses
D. be clear about the skills require to be demonstrated

21. Chenda, a student in Mr. Peo Virak‟s class, scored 78 marks on a reading test which has a mean of 80 and a standard deviation of 4. She scored 60 marks on the writing test which had a mean of 50 and a standard deviation of 3. Based on the

above information, in comparison to her peers, which statement provides the **most** accurate interpretation?

A. Chenda is better in reading than in writing.
B. Chenda is better in writing than in reading.
C. Chenda is below average in both subjects.
D. Chenda is close to average in both subjects.

22. After teaching four units from his course book, Mr. Peo Virak gives his students a test to measure their learning achievement. In this example, the **primary** purpose for conducting summative assessment is to

A. identify individual learning needs to plan classroom instruction.
B. motivate students to learn.
C. evaluate curriculum appropriateness.
D. determine the final grades for students.

23. Throughout instruction, Mr. Keo Ratana assesses how well his students are grasping the material. These assessments range from giving short quizzes, mid-term tests, written assignments to administering a semester examination. In order to improve the **validity** of this grading procedure, what advice should Mr. Peo Virak give to Mr. Keo Ratana?

A. consider students‟‟ class participation and their attendance before assigning a final grade.
B. consider students‟‟ performance in other subjects before assigning a final grade.
C. weight assessments according to their relative importance.
D. take into consideration each student‟‟s effort when calculating grades.

24. Ms. Meas Chakriya consults with Mr. Peo Virak for advice to effectively use her observations in recording her students‟‟ activities in the classroom. Which statement is **not** an appropriate recommendation when observing her students‟‟ behaviors?

A. make a record of the incident as soon after the observation as possible
B. maintain separate records of the factual description of the incident and her interpretation of the event
C. observe as many incidents in one long observation as possible
D. record both positive and negative behavioral incidents

25. Bora is a student in Mr. Keo Ratana‟‟s class. He receives a raw score of 12 items answered correctly out of a possible 15 on the vocabulary section of a test. This raw score equates to a percentile rank of 45. He is confused about how he could answer so many items correctly, but receive such a low percentile rank. He approaches Mr. Keo Ratana for a possible explanation. Which of the following is the **appropriate** explanation to offer to Bora?

A. “I don‟t know…there must be something wrong with the way the test is scored.”
B. “Although he answered 12 correctly, numerous students answered more than12 correctly.”
C. “Raw scores are purely criterion-referenced and percentile ranks are merely one form of norm-referenced scoring.”
D. “Raw scores are purely norm-referenced and percentile ranks are merely one form of criterion-referenced scoring.”
26. Prior to the semester examination, Mr. Keo Ratana reveals some information to his students. Which of Mr. Keo Ratana‟s action was **unethical**?

A. inform his students the exam contents to be covered.
B. inform his students the exam methods to be used.
C. show the actual exam paper to a small group of his low-achieving students.
D. tell his students the exam duration.

27. To achieve quality management of classroom assessments, Mr. Peo Virak advises his colleagues to be involved in all of the following **except**

A. quality assurance (concerning with quality of assessment by emphasising the assessment process).
B. quality teaching (dealing with the effectiveness of teaching in helping students undertake assessments successfully).
C. quality control (dealing with monitoring and, where necessary making adjustment to assessor judgments before results are finalised).
D. quality review (focusing on the review of the assessment results and processes in order to make recommendations for future improvement).

End of Test

Thank you for your kind help.

E. EVALUATION FORM FOR MODIFIED CAK

A language assessment literacy scale (attached), adapted and modified from Mertler (2003)'s *classroom assessment literacy inventory*, developed based on the seven standards listed by the *Standards for Teacher Competence in Educational Assessment of Students* (AFT, NCME, & NEA, 1990), one of the most recognised international standards in terms of teacher assessment literacy. This adapted and modified scale (consisting of 27 items), which adds two more standards (Tao, 2014), aims to investigate language assessment literacy knowledge of pre-service English teachers at a Turkish university who have just taken a compulsory course on English language assessment for an entire semester.

The scale contains a total of 27 items, each standard being addressed by three items each. The standards are:

1- Choosing Appropriate Assessment Methods
2- Developing Assessment Methods
3- Administering, Scoring, and Interpreting Assessment Results
4- Developing Valid Grading Procedures
5- Using Assessment Results for Decision Making
6- Recognising Unethical Assessment Practices
7- Keeping Accurate Records of Assessment Information
8- Ensuring Quality Management of Assessment Practices
9- Communicating Assessment Results

Before piloting the scale, you are kindly asked to evaluate each of the 27 items in the scale and provide expert view concerning whether the items are able to address and assess the related subcomponent (standard) of assessment literacy and have only one correct answer. You are also kindly asked to provide any kind of comments and remarks regarding the stems, options and scenarios in order to improve the items. In Table 1, please click the checkbox in the related column based on whether you think the item is appropriate or inappropriate. If you do not think the item is appropriate, please write your comments and remarks to improve the item in Table 2.

Thank you very much in advance for your kind support.

Table 1: Appropriateness

Item	Standard	Appropriateness	
		Appropriate	Inappropriate
1	Choosing Appropriate Assessment Methods	□	□
2	Developing Assessment Methods	□	□
3	Administering, Scoring, and Interpreting Assessment Results	□	□
4	Developing Valid Grading Procedures	□	□
5	Using Assessment Results for Decision Making	□	□
6	Recognising Unethical Assessment Practices	□	□
7	Keeping Accurate Records of Assessment Information	□	□
8	Ensuring Quality Management of Assessment Practices	□	□
9	Communicating Assessment Results	□	□
10	Choosing Appropriate Assessment Methods	□	□
11	Developing Assessment Methods	□	□
12	Administering, Scoring, and Interpreting Assessment Results	□	□
13	Developing Valid Grading Procedures	□	□
14	Using Assessment Results for Decision Making	□	□
15	Recognising Unethical Assessment Practices	□	□
16	Keeping Accurate Records of Assessment Information	□	□
17	Ensuring Quality Management of Assessment Practices	□	□
18	Communicating Assessment Results	□	□
19	Choosing Appropriate Assessment Methods	□	□
20	Developing Assessment Methods	□	□
21	Administering, Scoring, and Interpreting Assessment Results	□	□
22	Developing Valid Grading Procedures	□	□
23	Using Assessment Results for Decision Making	□	□
24	Recognising Unethical Assessment Practices	□	□
25	Keeping Accurate Records of Assessment Information	□	□
26	Ensuring Quality Management of Assessment Practices	□	□
27	Communicating Assessment Results	□	□

Table 2: Comments and Remarks

Item	Comments and Remarks
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
21	
22	
23	
24	
25	
26	
27	

References

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

Mertler, C. A. (2003). Pre-service versus in-service teachers‟ assessment literacy: Does classroom experience make a difference? *Paper presented at the annual meeting of the Mid- Western Educational Research Association*, October 15-18, Columbus, Ohio.

Tao, N. (2014). *Development and validation of classroom assessment literacy scales: English as a Foreign Language (EFL) instructors in a Cambodian Higher Education Setting* (Doctoral Dissertation). Victoria University.

F. MODIFIED CAK

PART 1

1- How old are you?

____

2- What is your gender?

A) Male
B) Female
C) Prefer not to say

3- What is your current CGPA on a 4.0 scale?

____

4- Have you ever been to a workshop or seminar in which the topic was only assessment?

A) Yes
B) No

5- Which of the following best describes your perception of the level of preparation for the overall job of being a classroom teacher?

A) Unprepared
B) Somewhat prepared
C) Prepared
D) Highly prepared

6- Which of the following best describes your perception of the level of preparation for assessing student performance?

A) Unprepared
B) Somewhat prepared
C) Prepared
D) Highly prepared

7- What is your e-mail address? (optional)

_________________________________@__________

PART 2

Scenario # 1

Mr. Alper Kaya, a first year English writing instructor, is aware of the fact that his students will be taking an end-of-the-term examination at the end of the course.

1- **Mr. Kaya wants to have information regarding his students' progress over time at the end of the semester so that he can reach conclusions about the progress the students have made in targeted writing skills. Which of the following assessment methods would be <u>the most</u> appropriate choice?**

A) Guided writing
B) Sentence building
C) Sequencing
D) Portfolio

2- **In order to grade his students' writing accurately and consistently, Mr. Kaya would be well advised to ----.**

A) identify criteria from the unit objectives and create a marking rubric
B) develop marking criteria after getting a feel for what students can do
C) consider student performance on similar types of tests
D) consult with experienced colleagues about a set of scoring criteria that has been used in the past

3- **Mr. Kaya wants to evaluate his students' essay-writing ability by providing specific feedback along a number of previously defined dimensions and descriptors. Which of the following would be <u>the best</u> tool to help him?**

A) Primary trait scoring rubric
B) Holistic scoring rubric
C) Objective scoring rubric
D) Analytic scoring rubric

4- Throughout the instruction, Mr. Kaya has given his students a number of assessments ranging from short quizzes following an introduction to a new topic, to administering an end-of-the-unit final exam. In order to improve the validity of this grading procedure, Mr. Kaya should ----.

A) make the grading scale the same for all assessments
B) consider students' prior performance before assigning a final grade
C) weight assessments according to their coverage
D) take into consideration each student's effort when calculating grades

5- At the end of each class period, Mr. Kaya asks his students several questions for formative assessment. The primary purpose for conducting formative assessment is to ----.

A) monitor how well the learning is progressing
B) determine the final grades for students
C) identify content for the final examination
D) evaluate curriculum appropriateness

6- Mr. Kaya was worried that his students would not perform well on the end-of-the-term examination. He did all of the following to help increase his students' scores. Which was unethical?

A) He instructed his students in strategies for taking tests.
B) He focussed in his instruction on concepts and skills to be covered on the test.
C) He planned and performed additional instruction for his low-scoring students.
D) He allowed students to practice with a small number of items from the actual test.

7- Mr. Kaya wants to keep the records of his assessment and reporting process. Which of the following is NOT one of the primary goals of keeping assessment records?

A) To help the teacher to use the same assessment tools in the assessment of future students
B) To enable the teacher and the student to reassess the teaching-learning relationship
C) To provide information and data regarding the future planning of the students' ongoing education
D) To facilitate the supply of information to administrators, parents and other stakeholders

8- To ensure the validity and reliability of his classroom assessment procedure, it is advised that Mr. Kaya should come together with his colleagues to discuss all of the following concepts. Which of the following is not a <u>primary</u> consideration in these discussions?

A) marking criteria
B) students' pieces of work
C) teaching techniques
D) assessment activities

9- In a routine conference with his students, Mr. Kaya is asked to explain the basis for assigning course grades. Mr Kaya should ----.

A) explain that the grading system was imposed by the school administrators
B) refer to the information that he presented to his students at the beginning of the course on the assessment process
C) encourage the students to do some research on what his grading could be based on through an out-of-class assignment
D) indicate that the grading system is imposed by the Ministry of Education

Scenario # 2

Ms. Zeynep Demir is an EFL teacher working in a high school setting. She has just finished teaching a unit on climate change and wishes to measure her students' understanding of this particular unit using a multiple-choice test where each item has only one correct option.

10- Based on Ms. Demir's goal, which of the following assessment strategies would be <u>the most</u> appropriate choice?

A) Using the test items included in the teacher's manual from the textbook she uses
B) Designing test items that are consistent with the content and skills specified in the course learning outcomes
C) Using the most popular test items that can be found on the Internet that cover the topic of climate change
D) Designing test items that cover a great deal of factual information on climate change

11- In constructing the multiple-choice test items, Ms. Demir should follow all of the following guidelines except ----.

A) repeat words or phrases in the questions and options
B) ensure that the correct answer is explicitly the best
C) avoid using phrases such as “none/all of the above” in the options
D) make all distractors plausible and attractive to the low ability test-taker

12- Ms. Demir decides to score the test using a 100% correct scale. Generally speaking, what is the proper interpretation of a student score of 85 on this scale?

A) The student answered 85% of the items on the test correctly.
B) The student knows 85% of the content covered by the instructional unit.
C) The student scored higher than 85% of other students who took this test.
D) The student scored lower than 85% of other students who took this test.

13- The multiple-choice test mentioned is the only student work that Ms. Demir grades for the current grading period. Therefore, grades are assigned only on the basis of this test. Which of the following is NOT a criticism of this practice?

A) The test, and thus the grades, reflects too narrow a curriculum focus.
B) The grades, since based on one test alone, are probably biased against some student groups.
C) Formal tests are more reliable than assessments in less standardized conditions such as homework.
D) Decisions like grades should be based on more than one piece of information.

14- Some of Ms. Demir’s students do not score well on the multiple-choice test. She decides that the next time she teaches this unit, she will begin by administering a pretest to check for students’ prerequisite knowledge. She will then adjust her instruction based on the pretest results. What type of information is Ms. Demir using?

A) Norm-referenced information
B) Criterion-referenced information
C) Both norm- and criterion-referenced information
D) Neither norm- nor criterion-referenced information

15- Based on their grades from last semester, Ms. Demir believes that some of her low-scoring students are brighter than their test scores indicate. Based on this knowledge, she decides to add some points to their test scores, thus raising their grades. Which of Ms. Demir's action was unethical?

A) Examining her students' previous performance
B) Adjusting grades in her course
C) Using classroom observations to adjust current grades
D) Adjusting low achieving students' grades

16- Ms. Demir understands that her classroom assessment records serve the following purposes except ----.

A) provide an overview of assessment methods developed
B) demonstrate diagnostic information regarding the students
C) show the extent of student progress throughout the instruction
D) inform administrative decision makers on various issues

17- To enhance the quality of a newly-developed multiple-choice test, Ms. Demir should do all of the following except ----.

A) pilot the test items with a small number of her past students to see how well each item performs
B) do a piloting study and make all necessary changes to the test items based on the information received
C) have all of her current students take the test twice and make a comparison of their scores
D) panel the test items through consultation with her colleagues who have assessment experience

18- During an individual conference, one student in Ms. Demir's class wants to know what it means that he scored in the 80th percentile in a multiple-choice test. Which of the following provides the best explanation for this student's score?

A) He got 80% of the items on the test correct.
B) He is likely to earn a grade of "B" in his class.
C) He has a score of 80 from the test on climate change.
D) He scored the same or better than 80% of his classmates.

Scneario # 3

Mr. Ahmet Kaplan is a senior EFL lecturer in a higher education setting. Experienced in issues of classroom assessment, Mr. Kaplan is often asked by his colleagues to respond to questions concerning best practices for evaluating student learning.

19- A colleague of Mr. Kaplan's asks what type of assessment is best to determine how well her students are able to apply what they have learned in class to a situation encountered in their everyday lives. Which of the following assessment concepts is <u>most</u> related to this situation?

A) diagnostic assessment
B) informal assessment
C) formative assessment
D) authentic assessment

20- A colleague of Mr. Kaplan's is constructing essay questions for a test to measure her students' critical thinking skills. She consults with Mr. Kaplan to see what concerns she would be aware of when constructing the questions. Which statement is <u>NOT</u> an appropriate recommendation when writing essay questions?

A) consider the relevance of the questions for a particular group of her students
B) avoid determining the amount of freedom of writing responses that will be accepted
C) indicate the time limits for the writing responses
D) be clear about the skills to be demonstrated

21- Ali, a student in Mr. Kaplan's class, scored 82 on a reading test which has a mean of 80 and a standard deviation of 4. He scored 60 on the writing test which had a mean of 50 and a standard deviation of 3. Based on the above information, in comparison to his peers, which statement provides <u>the most</u> accurate interpretation?

A) Ali is better in reading than in writing.
B) Ali is better in writing than in reading.
C) Ali is below average in both subjects.

D) Ali is close to average in both subjects.

22- Mr. Kaplan has made a number of assessment-related decisions throughout the semester. Which of the decisions least reflects students' achievement?

A) Reducing 5 points from a student's test grade for disruptive behaviour
B) Grading only the odd numbered items in a homework assignment
C) Using weekly quizzes and three major examinations to assign final student grades
D) Permitting students to redo their assignments when they need more opportunities to meet the standards for grades

23- After teaching four units from his course book, Mr. Kaplan gives his students a test to measure their learning achievement. In this example, the primary purpose for conducting summative assessment is to ----.

A) identify individual learning needs to plan instruction
B) motivate students to learn
C) evaluate curriculum appropriateness
D) determine grades for students

24- Mr. Deniz, one of Mr. Kaplan's colleagues, reveals some information to his students prior to the semester examination. Which of Mr. Deniz's actions was unethical?

A) Informing his students on the exam contents to be covered
B) Informing his students on the exam methods to be used
C) Showing his low-achieving students few items from the actual test
D) Telling his students the exam duration

25- A colleague of Mr. Kaplan's consults with him for advice to effectively use her observations in recording her students' activities in the classroom. Which statement is NOT an appropriate recommendation when observing students' behaviours?

A) Make a record of the incident as soon after the observation as possible
B) Maintain separate records of the factual description of each incident
C) Observe as many incidents in one long observation as possible
D) Record both positive and negative behavioural incidents

26- Mr. Kaplan conducts regular meetings with his colleagues to make a number of assessment-related decisions to comply with the fundamental principles in language assessment. Decisions regarding which of the following principles are most likely to negatively affect the quality of their assessment?

A) Reliability
B) Practicality
C) Validity
D) Authenticity

27- A student in Mr. Kaplan's class receives a raw score of 12 items answered correctly out of a possible score of 15 on the vocabulary section of a test. This raw score equates to a percentile rank of 45. He is confused about how he could answer so many items correctly, but receive such a low percentile rank. He approaches Mr. Kaplan for a possible explanation. Which of the following is the appropriate explanation to offer to the student?

A) "I don' know… there must be something wrong with the way the test is scored. I'll check immediately."
B) "Although you answered 12 correctly, numerous students in the class answered more than 12 correctly."
C) "Raw scores are purely criterion-referenced, but percentile ranks are merely one form of norm-referenced scoring."
D) "Raw scores are purely norm-referenced, but percentile ranks are merely one form of criterion-referenced scoring."

ANSWER KEY

1-D

2-A

3-D

4-C

5-A

6-D

7-A

8-C

9-B

10-B

11-A

12-A

13-C

14-B

15-D

16-B

17-C

18-D

19-D

20-B

21-B

22-A

23-D

24-C

25-C

26-B

27-B

G. TURKISH SUMMARY/TÜRKÇE ÖZET

DİLDE ÖLÇME-DEĞERLENDİRME OKURYAZALIĞINA YÖNELİK BİR ÖLÇEĞİN PSİKOMETRİK ÖZELLİKLERİNE DAİR BİR ARAŞTIRMA

Giriş

Bu bölümde eğitim bilimleri bağlamında ölçme-değerlendirme okuryazarlığı ve yabancı dil eğitimi bağlamında yabancı dilde ölçme-değerlendirme okuryazarlığı kavramları ve bu kavramların eğitim, öğretim ve öğrenim süreçleri içerisindeki yeri ve önemi konusunda kısaca bilgi verilmiştir. Bu bilgiler ışığında çalışmanın gerekçesi ve bu gerekçeye dayanarak oluşturulan araştırma hedeflerinden bahsedilmiştir. Bu hususlar devam eden paragraflarda özet olarak sunulmaktadır.

Ölçme-değerlendirme okuryazarlığı, eğitimsel ölçme-değerlendirme kavramının ve bununla ilgili becerilerin temel bir anlayışını ifade etmektedir ve ölçme değerlendirme okuryazarlığının öğretmenlerin sahip olması gereken temel bir beceri olduğu giderek daha fazla kabul görmektedir (Popham, 2009; Stiggins, 1991; Xu & Brown, 2016). Ölçme-değerlendirme konusunda donanımlı olan öğretmenlerin çeşitli bağlamlarda gerçekleştirilen ölçme-değerlendirme faaliyetlerinin ve politikalarının geçerliliği ve güvenilirliği hususunda daha isabetli ve çok yönlü kararlar alabileceklerine ilişkin geniş bir fikir birliği bulunmaktadır. Öte yandan, ölçme-değerlendirme okuryazarlığı konusunda zayıf olan öğretmenler ise geçerli ve güvenilir olmayan ölçme-değerlendirme uygulamaları yapma ve böylelikle hem öğrencileri, hem diğer paydaşları (okul idareciler, eğitim yetkilileri ve karar alıcıları ve ebeveynler vb.) hem de öğretim konusunda kendilerini yanlış yönlendirme riskiyle karşı karşıya kalabilmektedirler.

Ölçme-değerlendirme okuryazarlığı, öğretim ve öğrenim süreçlerinde önemli bir yere sahiptir. Ölçme-değerlendirme, öğrenimi etkilemekle kalmayıp onu şekillendirdiği için eğitimin ayrılmaz bir parçası olarak düşünülmektedir (White, 2009). Öğretmenler de, öğretimin çok önemli bir parçası olarak, çoğu zaman gerek sınıf içinde gerekse sınıf dışında resmi veya gayriresmî, geleneksel veya alternatif olmak üzere çok çeşitli ölçme-değerlendirme uygulamaları yapmak durumunda kalmaktadır. Bu uygulamalardan edindikleri bilgiyi ders içeriğinin uygun olup olmadığını belirlemek, öğrenme ve öğretme süreçlerini iyileştirmek, öğretimin ne kadar etkili olduğu konusunda fikir sahibi olmak ve öğrencilerin mevcut başarı durumlarına ve bir dersin öğrenme çıktıları / kazanımları bağlamında güçlü ve zayıf yanlarına dair onları bilgilendirmek gibi eğitim amaçları doğrultusunda kullanmaları beklenmektedir. Başka bir deyişle, öğretmenlerin eğitim-öğretim kapsamı içerisinde oldukça önemli bir yere sahip olan ölçme-değerlendirme sorumlulukları bulunmaktadır.

Ölçme-değerlendirme okuryazarlığının öğretmenlerin mesleki gelişimi içerisindeki önemiyle ilgili farkındalığın artmasıyla birlikte bu kavrama karşı akademik ilgi artmış ve ölçme-değerlendirmenin öğretmenlerin hem hizmet öncesi hem de hizmet içi eğitim programlarındaki yeri sorgulanmaya başlanmıştır (Alkharusi ve ark., 2011; Beziat & Coleman, 2015; Mertler, 2003; Xu & Brown, 2016). Bu akademik çalışmalarda öğretmenlerin ve öğretmen adaylarının ölçme-değerlendirme ile alakalı hem bilgileri hem de ilgileri yoklanmış ve eğitim içerisinde böylesi önemli bir yere sahip olan kavramla ilgili donanımlarının oldukça yetersiz olduğu gözlemlenmiştir. Bu yetersizlikler gerek sınıf düzeyinde ölçme-değerlendirme faaliyetleri gerekse büyük ölçekli ölçme-değerlendirme faaliyetleri için geçerli olmaktadır.

Eğitim bilimleri içerisinde gelişmiş olan ölçme-değerlendirme okuryazarlığı son yıllarda yabancı dil eğitimi alanına uyarlanmış ve bağlamda dilde ölçme-değerlendirme okuryazarlığı olarak ifade edilmeye başlanmıştır. Dilde ölçme-değerlendirme okuryazarlığının, özellikle kullanıldığı bağlam ve amaçlara göre değişen çeşitli tanımları yapılmıştır. Malone (2013) dilde ölçme-değerlendirme okuryazarlığını "dil eğitimcilerinin ölçme-değerlendirme tanımlarına olan aşinalığı ve bu bilginin genel olarak sınıf içi uygulamalara ve özel olarak dilde ölçme-değerlendirme meselelerine uygulanması" olarak tanımlamıştır (s. 9). Inbar-Lourie'ye (2008) göre dilde ölçme-değerlendirme bilgisi, ölçme-değerlendirme okuryazarlığı becerileri katmanlarına ek olarak dile özgü becerileri birleştiren ve dilde ölçme-değerlendirme okuryazarlığı şeklinde anılabilecek ayrı bir yapı teşkil eden bir temeldir (ss. 389-390). Başka bir deyişle, her ne kadar daha geniş eğitim bilimleri alanı içerisinde ortaya çıkmış olan ölçme-değerlendirme okuryazarlığı kavramından türetilmiş olsa da, dilde ölçme-değerlendirme kavramı dile özgü performansın kuramsallaştırılması ve ölçülmesi ile ilgili konuları da ihtiva etmektedir (Inbar-Lourie, 2017). Çeşitli araştırmalar, Stiggins'in (2010) eğitim alanında var olan "bol miktarda ölçme-değerlendirme bilgisizliğinin" dil eğitimi alanında da mevcut olduğunu ortaya koymuştur. Bu durum hem dünyanın çeşitli yerlerinde yapılan çalışmalar (Cheng, Rogers, & Hu, 2004; Diaz, Alarcon, & Ortiz, 2012; Lopez & Bernal, 2009; Tsagari & Vogt, 2017; Vogt & Tsagari, 2014; Volante & Fazio, 2007 vb.) hem de Türkiye'de yapılan çeşitli çalışmalarda (Hatipoğlu, 2015; Mede & Atay, 2017; Ölmezer-Öztürk & Aydın, 2018; Öz & Atay, 2017 vb.) gözlemlenmiştir. Ölçme-değerlendirme okuryazarlığının öğretim ve öğrenim açısından önemi ve öğretmenlerin bu konuda sergilemiş oldukları yetersizlikler öğretmenlerin, bu çalışma bağlamında İngiliz dili eğitimi öğretmenlerinin, dilde ölçme-değerlendirme okuryazarlığı seviyelerini ölçmenin mümkün olup olmadığının tartışılmasını gerekli hâle getirmektedir. Bu sebeple bu araştırma, (a) temel olarak dilde ölçme-değerlendirme okuryazarlığına yönelik geliştirilen bir ölçeğin (değiştirilmiş CAK ölçeği) psikometrik özelliklerinin incelenmesini amaçlamaktadır. Araştırma deseninin yapısına uygun olarak, bu araştırma hedefine ek iki araştırma hedefi daha ortaya konmuştur. Bunlardan bir tanesi, (b) Türkiye'de bulunan iki İngiliz Dili Eğitimi Bölümünde son sınıf öğrencisi olarak okuyan öğretmen adaylarının dilde ölçme-değerlendirme okuryazarlığı seviyelerini araştırmaktır. Son araştırma hedefi ise (c) örneklemde yer alan öğretmen adaylarının dilde ölçme-değerlendirme okuryazarlığı gelişimine katkıda bulunan faktörlerin (varsa) neler olduğunu incelemektir.

Literatür Tarama

Bu bölümde ilgili literatürün sırasıyla dilde ölçme-değerlendirmedeki temel hususlar, ölçme-değerlendirme ve öğretim ilişkisi, ölçme-değerlendirme okuryazarlığı ile dilde ölçme-değerlendirme okuryazarlığı konuları bağlamında bir incelemesi yapılmıştır.

Ölçme-değerlendirmenin temel hususları bağlamında öncelikle ölçme-değerlendirmedeki temel kavramlar ele alınmış; ölçme, ölçek, sınav ve değerlendirme gibi kavramlar detaylı bir şekilde incelenerek aralarındaki farklardan bahsedilmiştir. Amacına göre ölçme çeşitleri detaylı bir şekilde incelenmiş ve buna ek olarak ölçme türlerinde yer alan norm referanslı ölçme – ölçüt referanslı ölçme, özetleyici ölçme – süreç ölçme ve doğrudan ölçme – dolaylı ölçme gibi farklı dikotomilerin üzerinde durulmuştur. Geçerlik, güvenirlik, kullanışlılık ve puanlama olmak üzere ölçme-değerlendirmenin temel ilkeleri ele alınmıştır.

Takip eden iki kısımda ölçme-değerlendirme ile öğretim ilişkisi incelenmiş ve ölçme-değerlendirme okuryazarlığının tanımının kuramsal arka planı ele alınarak bu kavramın detaylı bir şekilde tanımlanmasına kullanılan kuramsal çerçevelerden bahsedilmiştir. Bu kapsamda, bu çalışmada kullanılan ölçeğin de kuramsal temelini oluşturan ve 1990 yılında Amerikalı Öğretmenler Federasyonu (American Federation of Teachers, Eğitimde Ölçme-değerlendirme Ulusal Konseyi (National Council on Measurement in Education) ve Ulusal Eğitim Sendikası (National Education Association) tarafından geliştirilmiş olan Eğitimde Ölçme-değerlendirmede Öğretmen Yeterlilikleri Standartları (AFT, NCME, & NEA, 1990) dokümanı hakkında bilgi verilmiştir. Bu doküman, özellikle sınıf içi ölçme-değerlendirme bağlamında öğretmenlerin sahip olması gereken nitelikleri ve kazanımları 7 ana standart hâlinde tanımlamaktadır.

Kalan kısımlarda ise öğretmenlerin ve eğitimcilerin ölçme-değerlendirme okuryazarlığını bu kuram çerçeveye dayalı olarak inceleyen ölçekler ele alınmış, ilgili araştırmalardan bahsedilmiş, ölçme-değerlendirme okuryazarlığının önemine değinilmiş ve son olarak dilde ölçme-değerlendirme özelinde gerek kuramsal çerçeve gerekse de bu konuda dünyada ve Türkiye'de yapılan çalışmalar irdelenmiştir.

Metodoloji

Amaç

Ölçme-değerlendirme uygulamalarının eğitimde oldukça önemli bir yeri olmasından dolayı ve öğretmenlerin eğitim-öğretimle ilgili zamanlarının en az üçte birlik bir dilimini ölçme-değerlendirmeyle ilgili uygulama ve faaliyetlere ayırmaları, onların ölçme-değerlendirme okuryazarlık düzeylerinin belirlenmesini zorunlu kılmaktadır. Ölçme-değerlendirme okuryazarlık düzeylerinin belirlenmesi, öğretmenlerin güçlü yanlarının ve zayıf yanlarının ortaya çıkarılarak ihtiyaçlarının tespit edilebilmesi ve böylelikle gerek hizmet öncesi eğitim gerekse de hizmet-içi eğitim programlarının şekillenmesi konusunda ciddi katkılar sağlayabilme potansiyeline sahiptir. Bu amaç doğrultusunda aşağıdakiler bu çalışmanın araştırma soruları olarak ortaya çıkmıştır:

Bu çalışmada cevap aranan araştırma soruları aşağıdaki gibidir:

1. *Standartlar* (Aft, NCME, & NEA, 1990) ve Mertler ve Campbell'ın (2005) ALI ölçeğinden uyarlanan değiştirilmiş CAK ölçeğinin psikometrik özellikleri nedir?
2. Türkiye'de yükseköğrenim bağlamındaki öğretmen adaylarının dilde ölçme-değerlendirme okuryazarlığı bilgi temeli ne düzeydedir?
3. Eğer varsa, dilde ölçme-değerlendirme okuryazarlığına katkıda bulunan faktörler nelerdir?

Bağlam ve Katılımcılar

Bu çalışmaya Türkiye'de faaliyet gösteren iki önemli devlet üniversitesinde okuyan dördüncü sınıf İngiliz Dili Eğitimi Bölümü öğrencileri katılmıştır. Toplam katılımcı sayısı 74'tür. Öğrencilerin tamamı çalışmaya katıldıkları tarih itibariyle en fazla bir dönem önce okudukları bölümlerde dilde ölçme-değerlendirme konusunda en az bir adet ders almışlardır. Katılımcılara değiştirilmiş CAK ölçeği, her biri yaklaşık 45 dakika süren ve gözetim altında ayrı oturumlarda uygulanmıştır. Ölçek; sıralama, yorgunluk ve kopya tarafından oluşabilecek bir yanlılığı engellemek adına üç farklı versiyon hâline getirilerek uygulanmıştır.

Veri Toplama Aracı

Bir eğitimcinin ölçme-değerlendirme okuryazarlık düzeyinin tam anlamıyla ölçülmesi o kişinin ölçme-değerlendirme konuları hakkındaki bilgi düzeyine ek olarak bu bilgiyi ölçme-değerlendirme uygulamaları yaparken ne ölçüde pratiğe dökebildiğine ölçmeyi gerektirmektedir. Özellikle uygulamaya geçirme başarısının incelenebilmesi için de hem sınıf içinde hem de sınıf dışında gözlemlenmesi ve aldığı ölçme-değerlendirme karar ve politikalarının öğretim, öğrenme ve hayat boyu öğrenme konuları açısından kısa, orta ve uzun vadede irdelenmesi gerekmektedir. Ancak ölçme-değerlendirme okuryazarlığının temelinde bilgi düzeyi yer almaktadır ve ölçme-değerlendirme okuryazarlığı bilgi düzeyiyle başladığı gibi, araştırmalar da bilgi düzeyi miktarı ile uygulamadaki nitelik arasında pozitif ve doğru bir ilişki olduğunu ortaya koymuştur (Bandura, 1997; Fishbein & Ajzen, 2010). Bu bulgudan hareketle ve pratiklik sebeplerinden dolayı bu çalışmada İngilizce dili eğitimi öğretmen adaylarının ölçme-değerlendirme okuryazarlığı bilgi düzeyinde incelenmesine karar verilmiş ve bu amaçla Tao (2014) tarafından *Standartlar* (AFT, NCME, & NEA, 1990) dokümanından ve ALI (Mertler & Camplbell, 2005) ölçeğinden uyarlanarak geliştirilen CAK ölçeğinin değiştirilmiş bir versiyonu kullanılmıştır. Orijinal ölçek *Standartlar* dokümanında belirtilen 7 standarda ek olarak kapsamlı bir literatür taraması sonucunda iki standart daha eklemiştir. Ölçülmek istenen toplamda 9 standart şunlardır:

1. Uygun ölçme-değerlendirme yöntemlerini seçebilme
2. Ölçme-değerlendirme yöntemleri geliştirebilme
3. Ölçme-değerlendirme sonuçlarını yönetebilme, puanlayabilme ve yorumlayabilme
4. Geçerli notlama ve derecelendirme prosedürleri geliştirebilme
5. Ölçme-değerlendirme sonuçlarını karar alma amacıyla kullanabilme
6. Etik olmayan ölçme-değerlendirme uygulamalarını tanıyabilme
7. Ölçme-değerlendirmeden elde edilen bilgi verileri doğru bir şekilde kayıt altına alabilme
8. Ölçme-değerlendirme uygulamalarının kalite yönetimini sağlayabilme
9. Ölçme-değerlendirme sonuçlarını öğrencilere, ailelere ve diğer paydaşlara aktarabilme

Üç senaryoya bağlı 27 maddeden oluşan ve her bir standardın üç madde ile temsil edildiği ölçeğin uyarlanması ve geliştirilmesi kapsamında, ölçek içeriği Türkiye'de bir İngiliz dili eğitimi sınıfı bağlamına uyarlanmış; ölçeğin senaryo ve maddelerinin içeriği ya revize edilmiş ya da tamamen değiştirilmiştir. Ayrıca, dil ve ifade tarzı bakımından da uygun görülen değişiklikler yapılmıştır. Bu süreçte yapılan değişikliklerle birlikte değiştirilmiş CAK ölçeği, ölçme-değerlendirme ve dilde ölçme-değerlendirme uzmanlarından oluşan bir panele bir değerlendirme protokolü ile gönderilmiştir. Onlardan alınan geri dönütlere göre bir sorunun iptal edilerek onun yerine aynı yapıya ve spesifikasyonlara dayanan yeni bir sorunun yazılması da dâhil olmak üzere başkaca değişiklikler de yapılarak son hâline getirilmiştir. Nihai versiyonda da üç senaryoya bağlı ve her bir standardın üçer soruyla temsil edildiği bir ölçek oluşturulmuştur.

Veri Analizi

Bir ölçme aracının psikometrik özelliklerinin araştırılmasında temel olarak geçerlik ve güvenirlik kavramları ön plana çıkmaktadır ve bu kavramların nicel olarak analiz edilmesinde Ölçme Kuramına farklı perspektifler getiren farklı yaklaşımlar bulunmaktadır. Bu yaklaşımlar temel olarak klasik yaklaşımlar (Klasik Test Kuramı) ve modern (Madde Tepki Kuramı) yaklaşımlar olarak ikiye ayrılmaktadır. Her iki grup içerisinde de çok çeşitli modellemeler bulunmaktadır. Aynı soruya farklı varsayım ve analitik tekniklerle cevap arayan bu iki yaklaşım büyük oranda benzer sonuçlara ulaşsa da (Fan, 1998), aralarındaki en büyük fark ölçme aracına tümsel ve grup bağımlı (Klasik Test Kuramı) veya madde bazında ve grup bağımsız (Madde Tepki Kuramı) bakmak olan her iki yaklaşımın da birbirlerine karşı avantajları ve dezavantajları bulunmaktadır. Bu sebeple, bir ölçme aracının psikometrik özelliklerinin incelenmesinde daha bütüncül bir resme ulaşabilmek için iki yaklaşımdan da faydalanılması gerekmektedir. Geçerlik ve güvenirlik kavramlarının nicel incelemesini bu yaklaşımlar güvenirlik katsayısı, standart hata katsayısı, madde güçlüğü, madde ayırt ediciliği gibi kavramlar üzerinden gerçekleştirmektedirler. Bu çalışmada hem klasik yaklaşım modeli (Klasik Test Teorisi) hem de verinin uygun olduğu modelin kullanıldığı Madde Tepki Kuramı modelleri kullanılmıştır. Madde Tepki Kuramı ailesinden bir parametreli lojistik model (1PL), Rasch modeli ve iki parametreli lojistik model (2PL) kullanılmıştır. Son olarak, ölçeğin boyutsallığını incelemek ve olası bileşenlerini tespit edebilmek amacıyla da Rasch Temel Bileşenler Analizi (TBA) gerçekleştirilmiştir. Bu analiz sonuçlarından birinci ve ikinci araştırma sorularının incelenmesinde faydalanılmıştır. Üçüncü araştırma sorusunun irdelenmesi için ise korelasyon analizleri yapılmıştır. Çalışmanın araştırma sonuçlarına yönelik olarak ne tür analizlerin kullanıldığı Tablo 1'de belirtilmektedir.

Tablo 1: Veri Analizinde Kullanılan Analitik Teknikler

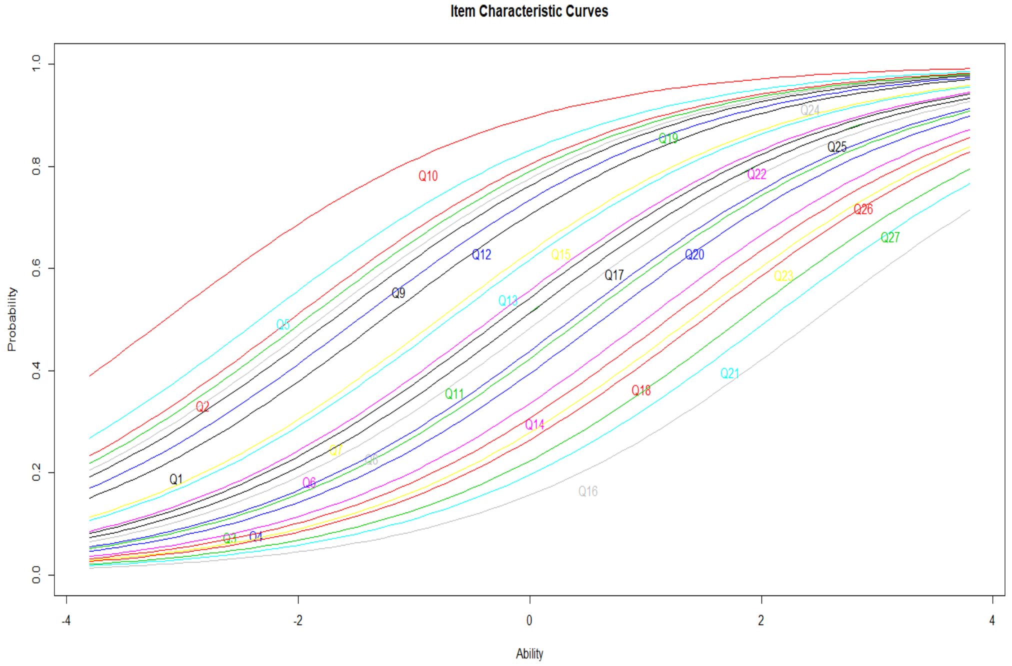
	Metot	Araştırma Sorusu
1	1PL	1 & 2
2	Rasch Analizi	1 & 2
3	Klasik Model	1 & 2
4	2PL	1 & 2
5	Rasch TBA	1
6	Betimleyici İstatistik	2
7	Korelasyon	3

Bulgular

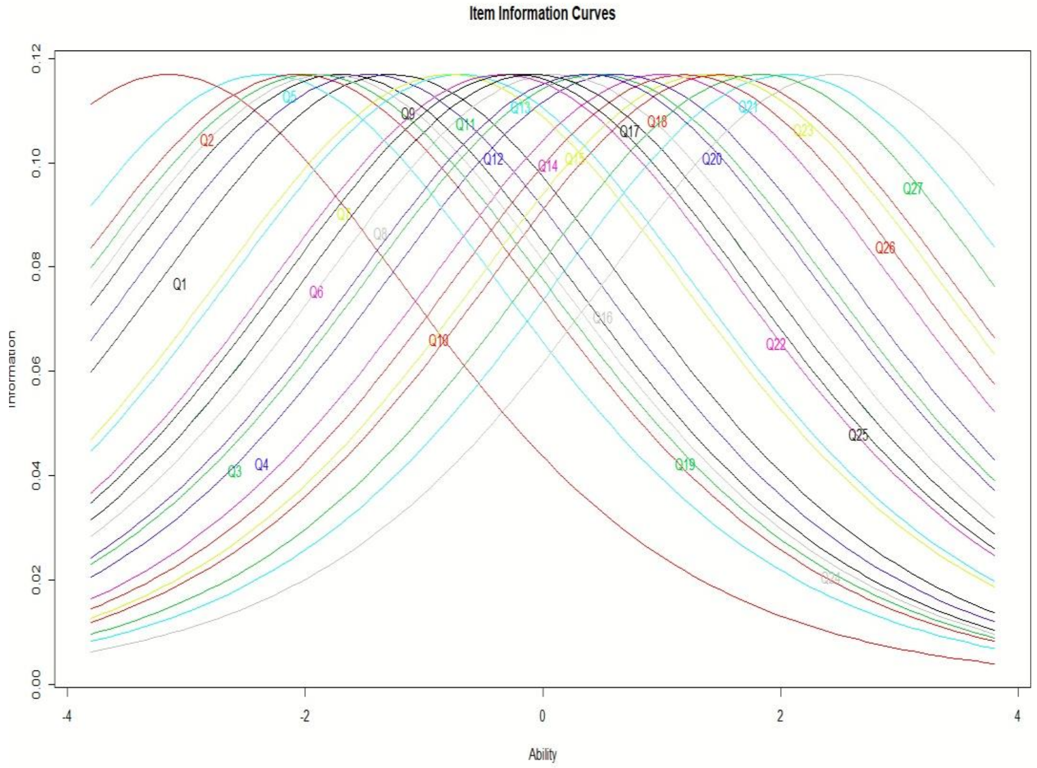
Bu bölümde 1PL modeli, Rasch modeli, klasik model ve 2PL modeli madde ve test analizlerinden elde edilen psikometrik bulgulara ek olarak, boyutsallığı ve bileşenleri inceleyen Rasch TBA analizinden ortaya çıkan bulgular rapor edilmiştir. En son olarak ise üçüncü araştırma sonucu bağlamında yapılan korelasyon analizi sonuçları paylaşılmıştır.

1PL Modeli

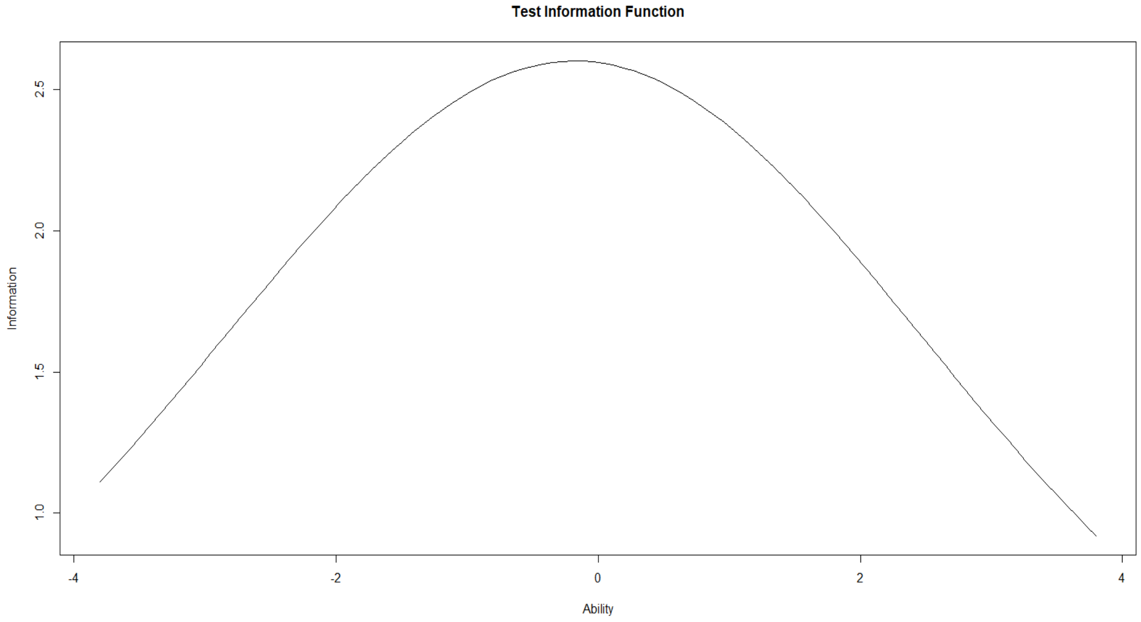
Model ve veri uyumluluğu kontrol edildikten sonra ilgili modele göre veri toplama süreci sonucunda elde edilen veriler incelenmiş ve her bir maddeye ait zorluk düzeyi (*b* parametresi) değerleri, maddelere ait madde karakteristik eğrileri ile madde bilgi eğrileri ve testin geneline ait test bilgi fonksiyonu eğrisi irdelenmiştir. Bu modelin -4.0 ve 4.0 yetenek aralığında toplam verinin %83,7'sini açıkladığı görülmüştür. Maddelere ait *b* parametresi değerleri -3.15 ve 2.47 arasında bir aralıkta çıkmıştır. Bu modelde sıfırın ortalama yetenek düzeyi olduğu varsayılmakta ve yetenek düzeyi eksende eksiye doğru daha düşük ve artıya doğru daha yükseğe doğru gitmektedir. Ortalama *b* değeri -0.20 olarak bulunmuştur. Sıfıra oldukça yakın olan bu değer, katılımcıların kuramsal olarak eksi sonsuz ve artı sonsuz aralığında bir yetenek düzeyine sahip olduğu varsayılan ancak pratiklik açısından katılımcıların -4.0 ve +4.0 arasındaki (yetenek düzeyini temsil eden) yatay eksene yerleştirildiği bu modelde, aşağı yukarı ortalama bir zorluk düzeyini ifade etmektedir. Madde karakteristik eğrileri, madde bilgi eğrileri ve test bilgi fonksiyonu 1, 2 ve 3 numaralı şekillerde sunulmaktadır.



Şekil 1: Tüm maddelere ait madde karakteristik eğrileri



Şekil 2: Tüm maddelere ait madde bilgi eğrileri



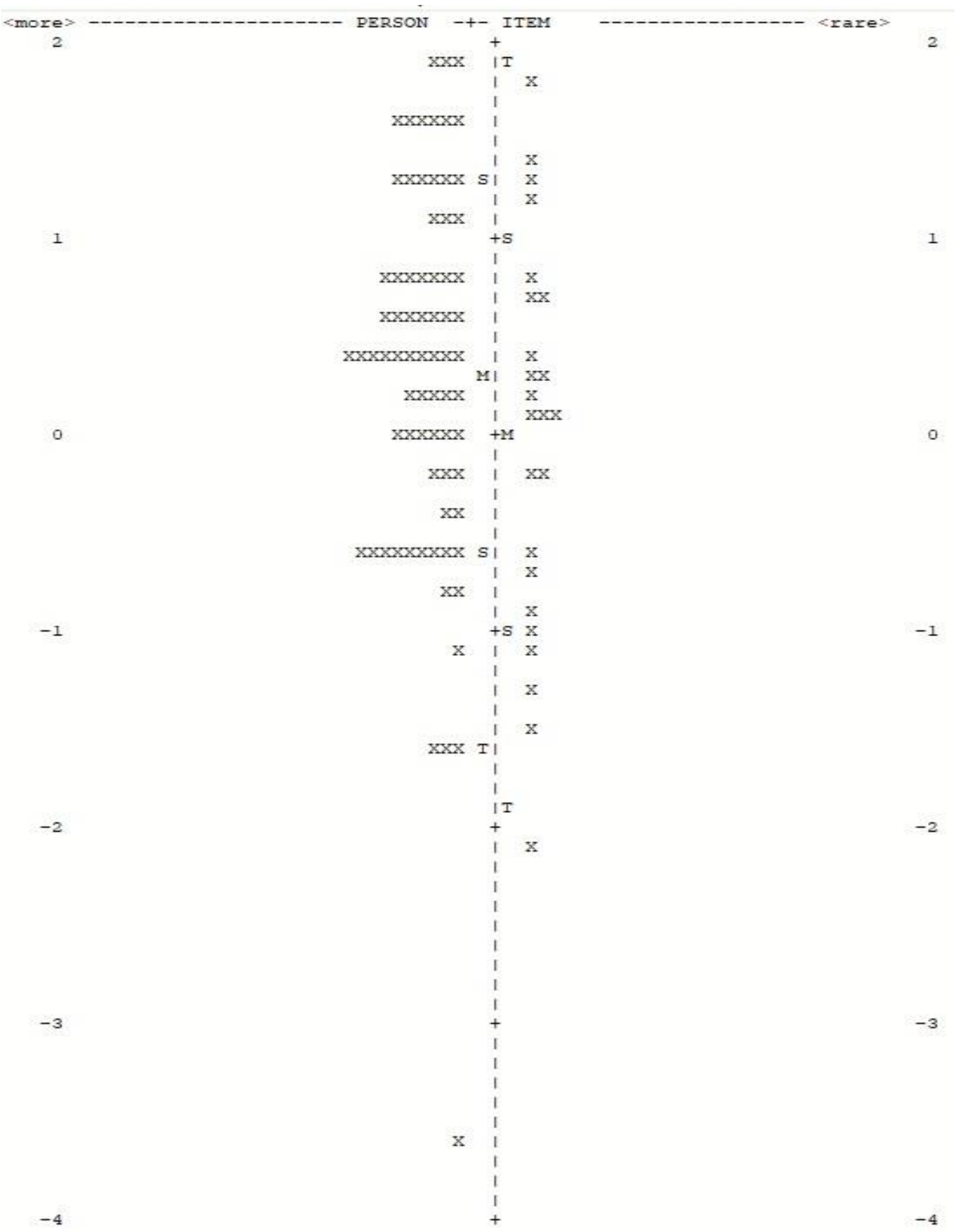
Şekil 3 Test Bilgi Fonksiyonu

Klasik Model

Veri, ikinci olarak klasik modele göre incelenmiş ve madde analizleri ile test analizleri sonuçlarına bakılmıştır. Bu analizde göze ilk çarpan detay ölçekte yer alan Q14, Q16 ve Q27 isimli maddelerin zayıf psikometrik sonuçlar ve iyi çalışmayan madde ayırt edicilik değerleri ürettiği gözlenmiştir. Seçenek ve içerik analizleri de yapıldıktan sonra bu maddelerin hatalı olduğu sonucuna varılmış ve ölçekten çıkartılmıştır. Bu aşamadan itibaren olan analizler bu üç madde hariç olmak üzere 24 madde üzerinden yapılmıştır. Klasik modeldeki analize göre ölçeğe dair veriler şu şekilde ortaya çıkmıştır: ölçekten elde edilen minimum ham skor 1, maksimum ham skor 20, ortalama ham skor 13,5'tur. Standart sapma değeri 4.10, varyans 16.82, çarpıklık -0.70, basıklık 0.25, KR20 değeri 0.73 ve ölçmenin standart hatası değeri 2.29 olarak bulunmuştur. Maddelerin güçlük düzeyi aralığı 0.22 ve 0.89, ayırt edicilik aralığı 0.19 ve 0.61 olarak bulunmuştur. Bu sonuçlara göre normal dağılım üreten ölçeğin kabul edilebilir düzeyde bir iç güvenirliği, oldukça iyi bir ayırt ediciliği ve ortalama bir güçlük düzeyi olduğu sonucu ortaya çıkmaktadır.

Rasch Analizi

Madde Tepki Kuramı ailesinden bir diğer model olan Rasch modeli, 1PL modeline oldukça benzer bir kuramsal arka plana sahip olmakla birlikte iki model arasındaki en temel fark, Rasch modelinin kalibrasyonu maddeler üzerinden yaparken 1PL modelinin kişiler üzerinden yapması ve Rasch modelinin logit değerlerini kullanarak madde-kişi haritası yoluyla madde-kişi eşleştirmelerini aynı ölçek üzerinde eşleştirmesidir. Ayrıca Rach modeli *infit* ve *outfit* modellemelerini yaparak her bir madde ve kişi için modele uyumluluk analizi yapmaktadır. 24 maddenin tamamı modele uygunluk değerleri içerisinde yer almıştır. Kişiler için 0.72 ve maddeler için 0.91 güvenirlik değerleriyle, bu analize göre de ölçeğin oldukça iyi güvenirlik katsayıları ürettiği gözlenmiştir. Maddelerin logit değer aralığı -2.11 ve 1.83'tür. Tüm maddelerin, dolayısıyla bir bütün olarak ölçeğin, ortalama logit değeri ise 0.01'dir. Diğer modellerle benzer şekilde bu model de ölçeğin ortalama bir güçlük düzeyine sahip olduğunu, farklı yetenek düzeyinde kişiler ve farklı yetenek düzeylerine hitap eden maddeler olduğunu ortaya koyarak ölçeğin nitelikli psikometrik özelliklere sahip olduğunu göstermiştir. Rasch analizinden elde edilen madde-kişi haritası (Wright haritası) Şekil 4'te sunulmaktadır:

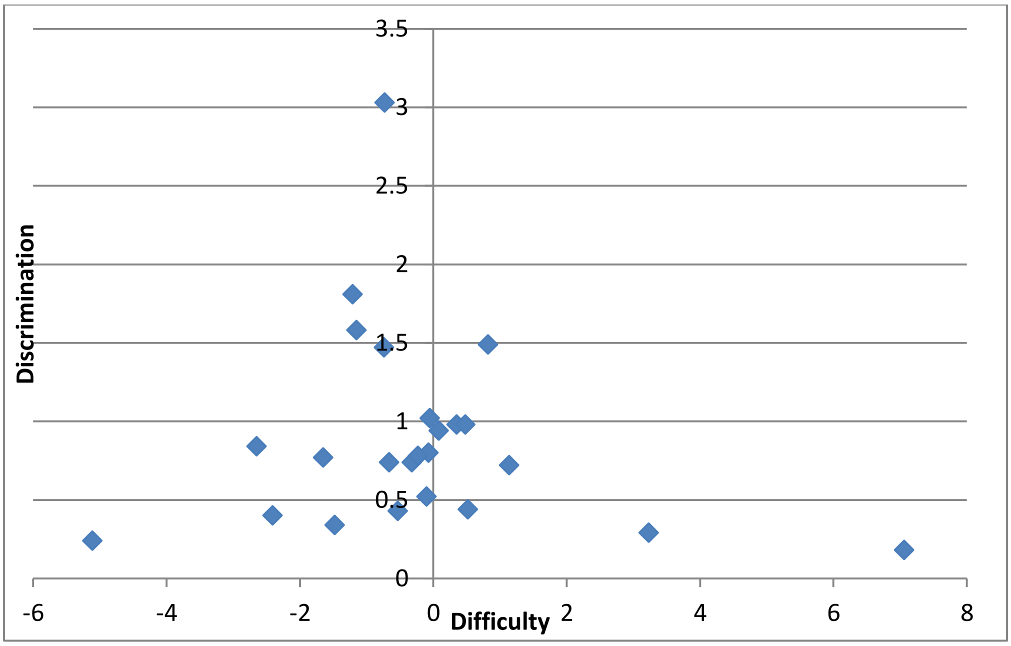


```
<more> --------------------- PERSON  -+- ITEM    ----------------- <rare>
    2                                 +                                   2
                                 XXX  |T
                                      |  X
                                      |
                              XXXXXX  |
                                      |
                                      |  X
                              XXXXXX S|  X
                                      |  X
                                 XXX  |
    1                                 +S                                  1
                                      |
                             XXXXXXX  |  X
                                      |  XX
                             XXXXXXX  |
                                      |
                          XXXXXXXXXX  |  X
                                     M|  XX
                               XXXXX  |  X
                                      |  XXX
    0                         XXXXXX  +M                                  0
                                      |
                                 XXX  |  XX
                                      |
                                  XX  |
                                      |
                           XXXXXXXXX S|  X
                                      |  X
                                  XX  |
                                      |  X
   -1                                 +S X                               -1
                                   X  |  X
                                      |
                                      |  X
                                      |
                                      |  X
                                 XXX T|
                                      |
                                      |
                                      |T
   -2                                 +                                  -2
                                      |  X
                                      |
                                      |
                                      |
                                      |
                                      |
                                      |
                                      |
                                      |
   -3                                 +                                  -3
                                      |
                                      |
                                      |
                                      |
                                      |
                                   X  |
                                      |
                                      |
                                      |
   -4                                 +                                  -4
```

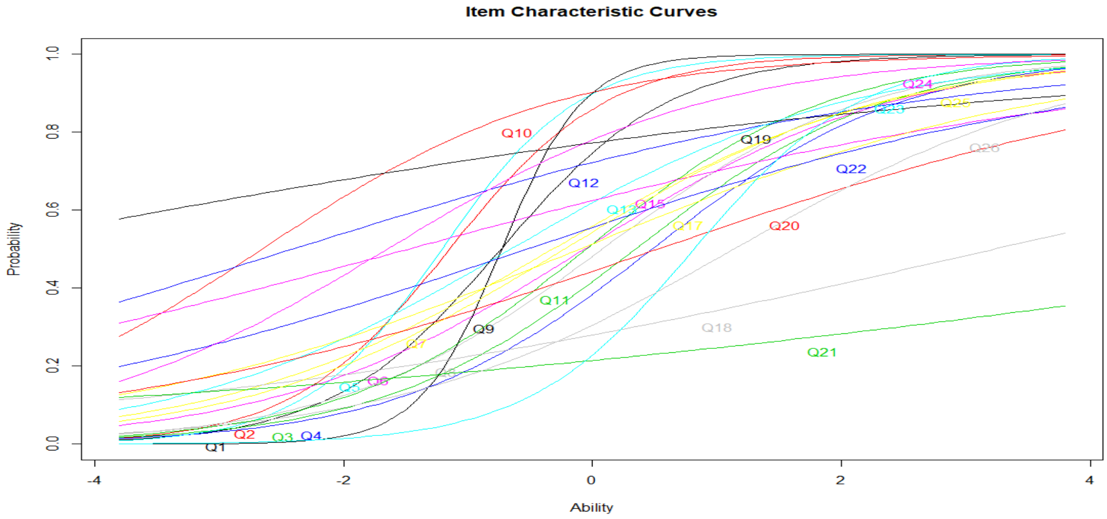
Şekil 4: Madde-kişi haritası (N=74)

2 PL Modeli

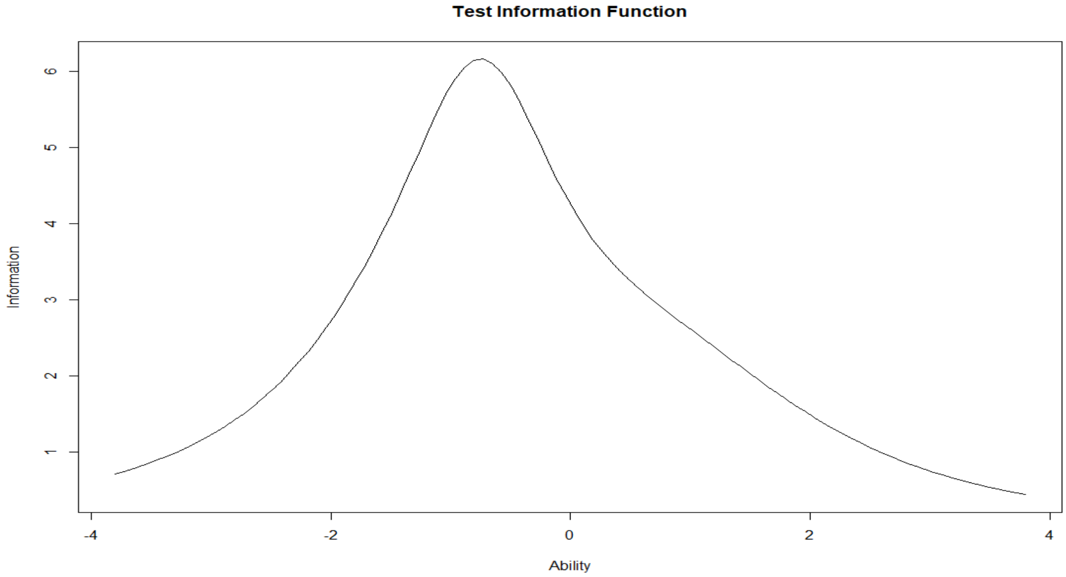
Madde ve test analizleri kapsamında son olarak 2PL modeli kullanılmıştır. Bu modelin kullanılma amacı, örtük özellik yaklaşımı bağlamında ayırt edicilik (a parametresi) bileşenini de inceleyen bir model kullanarak psikometrik özellikler bağlamında resmin tamamlayıcısı olacak bir teknik uygulama ihtiyacını gidermektir. Parça bazlı veri-model uyumluluğu test edildikten sonra analiz gerçekleştirilmiştir. Bu modelin -4.0 ve 4.0 yetenek düzeyi aralığında toplam verinin %91.46'sını açıkladığı gözlenmiştir. 1PL modele benzer şekilde, bu modele göre de ölçeğin ortalama bir güçlük düzeyine (-0.21) ve ortalama düzeyde bir ayırt edicilik değerine (*a*=0.87) sahip olduğu gözlenmiştir. Maddelere göre bakıldığında *b* parametresi aralığı -2.41 ve 7.06 iken *a* parametresi aralığı 0.18 ve 3.03'tür. Bu modele göre yapılan analizler de ölçeğin genel olarak geçerlik ve güvenirlik argümanlarını destekleyici kabul edilebilir psikometrik özelliklere sahip olduğunu göstermektedir. Bu analiz sonuçlarına göre ortaya çıkan güçlük-ayırt edicilik serpme diyagramı (Şekil 5), madde karakteristik eğrileri (Şekil 6) ve test bilgi fonksiyonu eğrisi (Şekil 7) görsel olarak sunulmuştur.



Şekil 5: Güçlük ve Ayırt Edicilik Serpme Diyagramı



Şekil 6: Madde Karakteristik Eğrileri



Şekil 7: Test Bilgi Fonksiyonu

Rasch PCA

Psikometrik özelliklerin analizi kapsamında son olarak elde edilen veriler ile Rasch PCA modeli kullanılarak boyutsallık analizi yapılmıştır. Bu analizin sonuçlarına göre ölçekte tek boyutluluk kriteri sağlanmış ve temel bileşenler irdelendiğinde, her ne kadar yapı tanımına göre dokuz ayrı yapıya (dokuz standart) yönelik olarak geliştirilmiş olsa da, ölçeğin genel olarak bir tanesi iki alt-bileşene sahip olmak üzere iki temel bileşene sahip olduğu gözlenmiştir. Bu durum, he ne kadar ölçeğin dayandığı kuramsal çerçeve ölçme-değerlendirme okuryazarlığını dokuz standarda ayırsa da her bir standardın birbiriyle etkileşim kurarak bir bütün oluşturduğu göz önüne alındığına doğal olarak değerlendirilmiştir. Bu analiz sonuçlarına göre ölçek iki temel bileşenden oluşmaktadır. Maddelerin içerik analizi yapıldığında, birinci bileşene ait olan maddelerin genel olarak ölçme-değerlendirme türleri arasındaki farklılıkları ayırt edebilme, öğrenci performansına nihai notlar verme, ölçme-değerlendirmede etik kaygılar ve ölçme-değerlendirmede geçerlik ve güvenirlik meseleleri konularına hitap ettikleri görülmüştür. İkinci bileşenin ise genel olarak ölçme-değerlendirme uygulamalarına yönelik uygun ölçekleri seçme ve geliştirme, öğrenci performansının puanlandırılması, ölçme-değerlendirme sonuçlarının kullanılması, yorumlanması ve paydaşlarla paylaşılması konularına hitap ettiği gözlenmiştir. Örneklem gurubunun ikinci bileşendeki performansının nispeten daha başarılı olduğu görülmüştür.

Korelasyon Analizi

Ölçme-değerlendirme okuryazarlığını etkileyen faktörleri tespit etmek amacıyla katılımcılardan elde edilen demografik ve arka olan verilerinin ve ölçekten elde edilen toplam ham skorların dâhil olduğu değişkenler arasındaki Pearson ve Spearman korelasyonlarına bakılmış ve istatistiki olarak anlamları ve pozitif korelasyonlar yalnızca "öğretmenlik mesleğine hazır olma algısı" derecesi ile "ölçme-değerlendirme uygulamaları yapmaya hazır olma algısı" derecesi arasında ve katılımcıların ağırlıklı genel not ortalaması ile ölçme-değerlendirme okuryazarlığı ölçeğinden elde ettikleri toplam ham puanlar arasında bulunmuştur.

Tartışma

Bu bölümde çalışmanın gerekçesi, metodolojisi ve araştırma soruları genel olarak gözden geçirilmiş ve buna göre ortaya çıkan çıkarımlar ve sonuçlar tartışılmıştır.

Araştırma Sorusu 1

Çalışmanın birinci araştırma hedefi, dil eğitimcilerinin ölçme-değerlendirme okuryazarlığını ölçme amacını taşıyan değiştirilmiş CAK ölçeğinin psikometrik özelliklerinin geçerlik ve güvenirlik temelinde incelenmesiyle ilgilenmiştir. Bu noktada hem bireysel olarak madde analizleri, hem de ölçeğin genel özelliklerine yönelik olarak test analizleri Ölçme Kuramına farklı bakış açıları getiren farklı yaklaşımların birlikte uygulanmasıyla yapılmıştır. Bu bağlamda kullanılan yaklaşımlar 1PL model, klasik model, Rasch modeli ve 2PL modeli olarak gerçekleşmiştir. Bu analizlere göre ölçekte bulunan üç soru zayıf psikometrik özellikler taşıdığı gerekçesiyle ölçekten çıkarılmıştır. Kalan 24 maddenin psikometrik özelliklerinin iyi düzeyde olduğu ve ölçeğin genel olarak iyi düzeyde geçerlik ve güvenirlik değerleri ürettiği rapor edilmiştir. Ölçek, Rasch TBA ile yapısal olarak da incelenmiş ve genel olarak ölçme-değerlendirme okuryazarlığı örtük özelliğini iki farklı bileşen etrafında ölçtüğü gözlenmiştir.

Araştırma Sorusu 2

Çalışmanın ikinci araştırma sorusu, Türkiye'de iki önde gelen devlet üniversitesinde dördüncü sınıf öğrencisi olarak okuyan ve en fazla bir dönem önce dilde ölçme-değerlendirme konusunda bir ders almış olan İngilizce öğretmen adaylarının dilde ölçme-değerlendirme okuryazarlığı (bilgi temelinde) ilgilenmiştir. Dilde ölçme-değerlendirme okuryazarlığını ölçen ve bu çalışmada psikometrik özellikleri incelenen ölçekten sağlanan veriler ışığında katılımcıların uluslararası tanınırlığı olan bir kuramsal çerçeve karşısında bilgi düzeyleri değerlendirilmiştir. Ölçekte kalan 24 madde üzerinden yapılan değerlendirmelerde adayların ortalama bir bilgi düzeyi sergiledikleri görülmüştür. Normal dağılım da göz önünde bulundurulduğunda, ortalama düzeydeki bir katılımcının ölçekte yer alan maddelerin yaklaşık olarak yarısını doğru cevaplayabildikleri gözlenmiştir. Katılımcılar arasında tüm maddeleri doğru cevaplayan olmazken, en yüksek ham skora sahip olan üç katılımcının doğru cevapladıkları toplam madde sayısı 20'dir. Kitlenin bariz bir şekilde en kolay bulduğu madde Q10 isimli maddedir. Bu madde, verilen senaryo bağlamında adaylardan duruma uygun ölçeği seçebilme kazanımını ölçmektedir. Kitlenin genel olarak en zor bulduğu madde ve en düşük başarı gösterdiği maddede ise (Q21) senaryo gereği bir sınıf içi değerlendirme uygulamasında kullanılan ölçeğe göre bir öğrencinin performansının yorumlanmasında birtakım matematiksel çıkarımlar yapabilme yetisini ölçmektedir. Kitle başarısıyla ilgili en dikkat çekici noktalardan biri, kitlenin öğrenci performansının yorumlanmasında persantil-ham skor ilişkiler, standart sapma-skor ilişkileri gibi birtakım matematiksel ve istatistiki hesaplamalar ve çıkarımlar yapmayı gerektiren maddelerde çok düşük başarı sergilemiş olmasıdır. Standartlar özelinde yapılan incelemeye göre ise, kitlenin en başarılı olduğu standartlar uygun ölçme metotlarının seçilmesi (Standart 1) ve etik olmayan ölçme-değerlendirme uygulamalarının tanınması (Standart 6) olmuştur. Diğer taraftan en başarısız olduğu standartlar ise ölçme sonuçlarının yönetimi, puanlaması ve yorumlanması (Standart 3) ve ölçme-değerlendirme uygulamalarının kalite yönetimi (Standart 8) olmuştur. Son olarak, kitle başarısı Rasch TBA analizinden elde edilen bileşenlere göre incelenmiş ve kitlenin ikinci temel bileşen konularında (amaca uygun ölçek seçme, öğrenci performansını puanlama, vb.) birinci temel bileşen konularına (amacına göre ölçme-değerlendirme türlerini ayırt edebilme, nihai notlandırma, geçerlik ve güvenirlik, vb.) kıyasla çok az bir farkla daha başarılı oldukları gözlemlenmiştir. Ancak neticede katılımcıların hiçbir standart özelinde topyekûn bir başarı sergilemediği, sergilenen başarının tüm standartlar ve bileşenler için aşağı yukarı ortalama düzeyde kaldığını ifade etmek mümkün olmaktadır. Bu durum, özellikle katılımcıların tamamının öğretmen olmak üzere olan bireyler oldukları ve kısa bir süre önce ölçme-değerlendirme konusunda bir ders almış oldukları düşünüldüğünde, ciddi kaygılara sebep olmaktadır.

Araştırma Sorusu 3

Çalışmanın üçüncü araştırma sorusu, dilde ölçme-değerlendirme okuryazarlığının gelişimine katkıda bulunan faktörler veya arka plan değişkenleriyle ilgilenmiştir. İlgili literatür şu ana kadar hâlihazırda olası faktörleri hizmet öncesi ölçme-değerlendirme eğitimi, deneyim, akademik nitelik, cinsiyet, mesleki gelişim, öğretmenlikte geçen süre, sınıf mevcudu ve öğrenci olarak ölçme-değerlendirme deneyimi olarak bulmuştur. Bu çalışmanın araştırma deseni gereği, katılımcı homojenliği göz önünde bulundurulduğunda burada sayılan olası faktörlerden çok az bir kısmına bakmak mümkün olmuştur ve elde edilen değişkenler içerisinde ölçekten alınan toplam ham skor ile yalnızca ağırlıklı genel not ortalamasının anlamlı ve pozitif bir ilişkisi olduğu gözlenmiştir. Başka bir deyişle, genel olarak bir katılımcının ağırlıklı genel not ortalaması ne kadar yüksekse bu ölçekten edindiği ham skor da o kadar yüksek olmuştur.

Çıkarımlar

Elde edilen bulgular doğrultusunda psikometrik ve pedagojik birtakım çıkarımlarda bulunmak mümkün olmuştur. Çalışmanın birinci ve ana araştırma hedefi doğrultusunda gerçekleştirilen ve Ölçme Kuramına farklı bakış açıları sunan psikometrik analizler öncelikle değiştirilmiş CAK ölçeğinin ilk uygulama versiyonunda bulunan üç maddenin problemli maddeler olduğunu ortaya koymuştur. Bu ölçeği ileride kullanmayı düşünebilecek olan araştırmacıların, ilgili spesifikasyonlara göre bu maddeleri yeniden üretme aşamasında bu hususu göz önünde bulundurmaları tavsiye edilmektedir. Ancak, ölçekte geri kalan maddelerin tümünün, bu çalışma sonucunda elde edilen psikometrik özelliklere göre, gelecekte aynı amaçla kullanılmak üzere kopyalanabileceği veya uyarlanabileceği değerlendirilmektedir. Ancak ölçme-değerlendirme okuryazarlığı kavramı bağlam bağımlı olduğu için (Inbar-Lourie, 2008) yapılacak her replikasyon veya adaptasyonun ihtiyaçları göz önünde bulundurarak şekillendirilmesi önerilmektedir. Replikasyon ve/veya adaptasyonların öğretmen eğitimcileri tarafından gerek öğretim gerek öğrenim bağlamında ihtiyaç analizi ve revizyon amaçlarıyla; öğretmen eğitimi karar alıcılarının da sınıf düzeyinde ve/veya program düzeyinde alınan kararları bilgilendirme amacıyla kullanabilecekleri değerlendirilmektedir.

Pedagojik bağlamda ise, ilgili literatürdeki çalışmaların hâlihazırda ortaya koyduğu gibi bu çalışmanın da ölçme-değerlendirme okuryazarlığının eğitim alanı içerisinde çok önemli bir yeri olmasına ve hatta ölçme-değerlendirme ve eğitimin birbirlerini sürekli besleyen bir bütünün parçaları olmalarına rağmen öğretmen adaylarının (ve öğretmenlerin) bu konudaki bilgilerinin istenen ve beklenen düzeylerde olmadığı çıkarımı yapılmaktadır. Elde edilen sonuçlar; örneklem gurubunun dördüncü sınıf öğrencileri olmaları ve tüm program süresince olan ders yüklerinin (ölçme-değerlendirme dâhil) büyük bir bölümünü tamamlamış olmaları göz önünde bulundurulduğunda, durumu daha ciddi kılmaktadır. Dolayısıyla, örneklem grubunun değiştirilmiş CAK ölçeğinde sergilemiş olduğu performans, programlarının onları ölçme-değerlendirme konularına ne düzeyde hazırladıklarını göstermesi açısında önemli bulunmuştur.

Sonuç

Eğitim bilimlerinde ölçme-değerlendirme okuryazarlığının alan içerisinde kendisine daha fazla yer bulduğu gibi, dilde ölçme-değerlendirme okuryazarlığı da dil eğitimi içerisinde hem dünya çapında hem de Türkiye'de gittikçe daha fazla ilgi görmektedir. Bu değişim sürecinde hem öğretmenler hem de öğretmen yetiştirme programları ön planda yer almaktadır. Bu çalışmada dil eğitimcilerinin dilde ölçme-değerlendirme okuryazarlığı bilgi temelini ölçekte kullanılan değiştirilmiş CAK ölçeğinin psikometrik özellikleri incelenmiştir. Psikometrik inceleme kapsamında 1PL ve 2PL Madde Tepki Kuramı modelleri, Rasch modeli ve klasik model kullanılmıştır. Ayrıca ölçeğin yapısının boyutsallığının ve temel bileşenlerinin incelenmesi için de Rasch TBA kullanılmıştır. Son olarak ise kişi arka plan değişkenleri ile ölçek performansı arasındaki ilişkiler istatistiki olarak incelenmiştir.

Sonuçlar 27 maddelik ölçeğin üç maddesinin problemli olduğunu, geri kalan maddelerle birlikte ölçeğin ölçme amacı bağlamında geçerlik ve güvenirliğin varlığına işaret eden kanıtlar olduğunu ve ölçeğin aynı amaçla replikasyonunun ve adaptasyonunun yapılabileceğini göstermiştir. Örneklem grubunun başarı durumuyla ilgili olarak ise ölçme-değerlendirme okuryazarlığı konusunda ortalama bir başarı sergiledikleri ancak bu ortalama başarının kendilerinden istenen ve beklenen düzeylerin çok altında olduğunu ortaya koymuştur. Korelasyon analizleri ağırlıklı genel akademik not ortalamasının ölçekten elde edilen başarıyla anlamlı ve pozitif ilişkisi olan tek değişken olduğuna işaret etmiştir. Öğretmen eğitimcileri, öğretmen yetiştirme programları, öğretmen adayları, öğretmenler ve araştırmacılar olmak üzere tüm olası paydaşların bu araştırma sonuçlarına temkinli ve dikkatli bir şekilde yaklaşmaları tavsiye edilmekte birlikte, dilde ölçme-değerlendirmeyi bilgi düzeyinde ölçen bu ölçeği kullanmayı tercih edebilecekleri değerlendirilmektedir.

H. TEZ İZİN FORMU / THESIS PERMISSION FORM

TEZ İZİN FORMU / THESIS PERMISSION FORM

ENSTİTÜ / INSTITUTE

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences □

Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences ■

Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics □

Enformatik Enstitüsü / Graduate School of Informatics □

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences □

YAZARIN / AUTHOR

Soyadı / Surname : Yılmaz
Adı / Name : Fahri
Bölümü / Department : İngiliz Dili Eğitimi

TEZİN ADI / TITLE OF THE THESIS (**İngilizce** / English) : An Investigation of the Psychometric Properties of a Language Assessment Literacy Measure

TEZİN TÜRÜ / DEGREE: **Yüksek Lisans** / Master ■ **Doktora** / PhD □

1. **Tezin tamamı dünya çapında erişime açılacaktır. /** Release the entire work immediately for access worldwide. ■

2. **Tez iki yıl süreyle erişime kapalı olacaktır.** / Secure the entire work for patent and/or proprietary purposes for a period of **two years. *** □

3. **Tez altı ay süreyle erişime kapalı olacaktır.** / Secure the entire work for period of **six months**. ***** □

***** *Enstitü Yönetim Kurulu kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir.*
A copy of the decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.

Yazarın imzası / Signature .... **Tarih** / Date: 23/07/2020