



HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Foreign Language Education

English Language Teaching

CEFR ORIENTED TESTING AND ASSESSMENT PRACTICES IN NON-FORMAL
ENGLISH LANGUAGE SCHOOLS IN TURKEY

Nurdan KAVAKLI

Ph.D. Dissertation

Ankara, 2018



With leadership, research, innovation, high quality education and
change,

To the leading edge... Toward being the best...



HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Foreign Language Education

English Language Teaching

CEFR ORIENTED TESTING AND ASSESSMENT PRACTICES IN NON-FORMAL
ENGLISH LANGUAGE SCHOOLS IN TURKEY

TÜRKİYE'DEKİ İNGİLİZCE KURSLARINDA AVRUPA DİLLERİ ÖĞRETİMİ
ORTAK ÇERÇEVE PROGRAMI ODAKLI ÖLÇME VE DEĞERLENDİRME
UYGULAMALARI

Nurdan KAVAKLI

Ph.D. Dissertation

Ankara, 2018

Acceptance and Approval

To the Graduate School of Educational Sciences,
This dissertation entitled "CEFR Oriented Testing and Assessment Practices in Non-Formal English Language Schools in Turkey" has been approved as a dissertation for the Degree of **Ph.D.** in the **Program of English Language Teaching** by the members of the Examining Committee.

Chair

Prof. Dr. Mehmet DEMİREZEN

Signature

Member (Supervisor)

Prof. Dr. İsmail Hakkı MİRİCİ

Signature

Member

Assoc. Prof. Dr. Nuray ALAGÖZLÜ

Signature

Member

Assoc. Prof. Dr. Hacer Hande UYSAL

Signature

Member

Assist. Prof. Dr. Hatice ERGÜL

Signature

This is to certify that this dissertation has been approved by the aforementioned examining committee members on 16/01/2018 in accordance with the relevant articles of the Rules and Regulations of Hacettepe University Graduate School of Educational Sciences, and was accepted as a **Ph.D. Dissertation** in the **Program of English Language Teaching** by the Board of Directors of the Graduate School of Educational Sciences on/...../.....

Prof. Dr. Ali Ekber ŞAHİN
Director of Graduate School of Educational Sciences

Abstract

As an offspring of the negotiation on the accession of Turkey to the European Union, the Framework has been adopted as a basis for language education policies with related implementations in the field of language testing and assessment. Although there is a growing body of research on the CEFR and language teaching, testing and assessment in formal educational settings, there is a perceived gap in the literature regarding the CEFR oriented testing and assessment practices of English language schools serving as non-formal educational settings. Considering this research gap, the prevalent aim of this study is to scrutinize the appropriateness of the current testing and assessment practices of English language schools rendering non-formal education to some European guidelines together with the Framework. Grounded upon a mixed methods research design, this study embraces quantitative data gathered from English language teachers who are also test (-item) developers at 3 private institutions rendering non-formal English language education, which are renowned for quality with the highest course attendee capacity and branches in Turkey together with the qualitative data gathered from their directors and that of the Association of Private Educational Institutions and Study Centers in Turkey (ÖZ-KUR-DER). Accordingly, the results have yielded that English language schools do not apply European guidelines in language testing and assessment thoroughly as the Framework is not adequately covered in related practices. This study has several implications for research on the development of the CEFR oriented language testing and assessment practices, and feeds into the growing body of research on the utilization of the European standards in non-formal educational settings. To sum up, in this study, the current language testing and assessment practices in non-formal educational settings, as the arteries of Turkish education economy, have been discussed to improve the quality by the exploitation of the CEFR.

Keywords: language testing, assessment, non-formal education, standardized tests, EFL, CEFR.

Öz

Türkiye'nin Avrupa Birliği müzakere süreci kapsamında benimsediği ve yürütmekte olduğu dil eğitim politikalarının dayanağı olan ADOÇEP, eğitimde ölçme ve değerlendirme alanında da uygulanmaktadır. Yapılan alanyazın taramasına göre, yaygın eğitim kapsamında ölçme ve değerlendirme uygulamalarını irdeleyen çalışmaların yanı sıra, Türkiye'de yaygın eğitim hizmeti veren İngilizce kurslarının ölçme ve değerlendirme uygulamaları açısından daha önce yapılmış bir çalışmaya rastlanmamıştır. Buna göre, bu çalışmanın temel amacı Türkiye'de yaygın eğitim hizmeti veren İngilizce kurslarının ölçme ve değerlendirme uygulamalarının ADOÇEP ile birlikte birtakım Avrupa ölçme ve değerlendirme ölçütlerine uygunluğunu irdelemektir. Çalışmada temel alınan Avrupa ölçme ve değerlendirme ölçütleri, Avrupa Dil Ölçme ve Değerlendirme Birliği, Avrupa Dil Testleri Uygulayıcılar Birliği, Uluslararası Ölçme Birliği ve Avrupa Dil Eğitim Değerlendirme Birliği tarafından öngörülen uygulama esaslarıdır. Metodolojik açıdan bu doktora tezi, karma yöntemle dayalıdır. Nicel veri, 5 seçenekli Likert tipi ölçek yoluyla, Türkiye'de yaygın eğitim hizmeti veren, ülke genelinde kalitesiyle bilinen 3 İngilizce kursunun belirlenmesiyle, bu kurumlarda sınav hazırlayıcı olarak çalışan İngilizce öğretmenlerinin görüşleri alınarak toplanmıştır. Nitel veri ise İngilizce kurslarının ve 'Tüm Özel Öğretim Kursları, Hizmet İçi Eğitim Merkezleri, Dershaneler ve Etüt Eğitim Merkezleri Birliği Derneği (ÖZ-KUR-DER)'nin yöneticileri ile 6 soru üzerinden yapılan yarı-yapılandırılmış görüşme yöntemiyle elde edilmiştir. Bu çalışmanın bulguları İngilizce kurslarında çalışan öğretmenlerin ilgili ölçütler hakkında yeterince bilgi sahibi olmadığını göstermiştir. Ayrıca, kurum yöneticilerinden toplanan görüşler ise ADOÇEP'in ölçme ve değerlendirme konusunda yeterli düzeyde uygulamaya konulmadığını ortaya çıkarmıştır. Bu çalışma, yaygın eğitim hizmeti veren İngilizce kurslarının ölçme ve değerlendirme çalışmalarının Avrupa standartları çerçevesinde geliştirilmesi konusunda çeşitli çıkarımlar da sunmaktadır. Özetle, bu çalışmada Türkiye'de yaygın eğitim yoluyla İngilizcenin öğretiminde benimsenen ölçme ve değerlendirme uygulamalarının ADOÇEP kapsamında gelişimi tartışılmıştır.

Anahtar sözcükler: ölçme ve değerlendirme, yaygın eğitim, ölçünleştirilmiş sınav, ADOÇEP, İngilizcenin yabancı dil olarak öğretimi.

Acknowledgements

First, I heartily thank my dear supervisor, Prof. Dr. İsmail Hakkı MİRİCİ, whose encouragement, guidance and support have allowed me to grow as a researcher. You have always been a tremendous mentor and an academic father for me with your priceless advice both on my researches as well as on my career plans.

Secondly, I am thankful to the committee members, Prof. Dr. Mehmet DEMİREZEN, Assoc. Prof. Dr. Nuray ALAGÖZLÜ, Assoc. Prof. Dr. Hacer Hande UYSAL and Assist. Prof. Dr. Hatice ERGÜL. I would like to express my sincere gratitude to them for their guidance and invaluable comments.

Thirdly, I am very grateful to TÜBİTAK (The Scientific and Technological Research Council of Turkey) for their support lasting for five years under the name of 2211-Graduate Scholarship Program, and to ÖYP for the funding offered since 2012.

To add more, I would like to thank my family and my friends: to my parents, my brother and my second family in Ankara (my aunt, her husband and their little daughter) for their never-ending support throughout writing this dissertation and my life in general; and to my dear friends from 'OLDIES but GOLDIES' for their incentives while I am striving towards my goal. Haters gonna hate, you know!

My special thanks go to my dear colleague Assist. Prof. Dr. Emrah DOLGUNSÖZ for his continuous support, love, encouragement and precious suggestions, to my bosom friend Res. Assist. Sezen ARSLAN for always being at my elbow, and to my partner-in-crime Özge AKARSU for standing by me since 2006.

Finally, I would like to extend my heartfelt appreciation to the participants in the study, recruited from anonymous private institutions rendering non-formal English language education, and to their dearworth directors together with that of ÖZ-KURDER, Hilmi ALPAN, whose contributions made this research possible.

Table of Contents

Abstract.....	ii
Öz.....	iii
Acknowledgements.....	iv
List of Tables	vii
Symbols and Abbreviations	viii
Chapter 1 Introduction	1
Background of the Study: A Personal Perspective	1
Motivation for the Study.....	2
Research Objectives	3
Significance of the Study.....	4
Assumptions.....	5
Limitations	5
Definitions of the Key Terms	6
The Organization of the Dissertation	7
Chapter 2.....	9
The European Standards of Language Testing and Assessment.....	9
Introduction.....	9
The CEFR: Origins, Content and Development.....	9
The EALTA: Mission, Principles and Considerations.....	23
The ALTE: Objectives, Standards and Resources	28
ILTA: The Objectives, Specifications and Resources.....	36
The AEA- Europe: Purpose, Core Elements and Guiding Principles.....	42
Chapter Summary	49
Chapter 3.....	51
The National Standards of Language Testing and Assessment in Turkey	51
Introduction.....	51
The Turkish National Education System.....	51
Language Testing and Assessment in Formal Educational Settings.....	53
Language Testing and Assessment in Non-Formal Educational Settings	55
Other Relevant Units of Language Testing and Assessment	57
Chapter Summary	58
Chapter 4.....	59
Methodology	59
Introduction.....	59

Research Design.....	59
Participants and Setting.....	60
Instruments.....	65
Procedure.....	69
Data Analysis	70
Ethical Considerations.....	73
Chapter 5.....	74
Findings and Results	74
Introduction.....	74
Results of the Data Analysis.....	74
Chapter 6.....	197
Conclusion and Discussion.....	197
An Overview of the Study	197
Discussion of the Results	199
Pedagogical Implications.....	207
Suggestions for Further Studies	210
Conclusion.....	211
References	216
APPENDIX-A: Questionnaire on the European Standards for Establishing Quality Profiles in Language Examinations.....	230
APPENDIX-B: Semi-Structured Interview Forms conducted with the Directors of Selected Private Institutions and that of ÖZ-KUR-DER (Original in Turkish)	236
APPENDIX-C: Semi-Structured Interview Forms conducted with the Directors of Selected Private Institutions and that of ÖZ-KUR-DER (Translated into English)	237
APPENDIX-D: Ethics Committee Approval.....	238
APPENDIX-E: Declaration of Ethical Conduct.....	239
APPENDIX-F: Dissertation Originality Report.....	240

List of Tables

Table 1. Overall Demographic Information of the Participants.....	61
Table 2. Demographic Information on the Institution A.....	62
Table 3. Demographic Information on the Institution B.....	64
Table 4. Demographic Information on the Institution C.....	65
Table 5. Reliability Co-Efficiency of the Data Collection Instrument.....	66
Table 6. An Outline of the Minimum Standards in the Questionnaire.....	67
Table 7. Questionnaire Items by the Guidelines of the EALTA.....	75
Table 8. The Exploitation of the EALTA Guidelines by Selected Private Institutions.....	76
Table 9. The Implementation of the EALTA Guidelines by Selected Private Institutions.....	78
Table 10. Questionnaire Items by the ALTE Code of Practice.....	95
Table 11. The Exploitation of the ALTE Code of Practice by Selected Private Institutions.....	99
Table 12. The Implementation of the ALTE Code of Practice by Selected Private Institutions.....	104
Table 13. Questionnaire Items by the ILTA Guidelines for Practice.....	144
Table 14. The Exploitation of the ILTA Guidelines for Practice by Selected Private Institutions.....	145
Table 15. The Implementation of the ILTA Guidelines for Practice by Selected Private Institutions.....	146
Table 16. Questionnaire Items of the European Framework of Standards for Educational Assessment by the AEA- Europe.....	157
Table 17. The Exploitation of the AEA- Europe's Framework of Standards by Selected Private Institutions.....	159
Table 18. The Implementation of the AEA- Europe's Framework of Standards by Selected Private Institutions.....	162
Table 19. The Utilization of the European Guidelines in Testing and Assessment Practices by Selected Private Institutions.....	187

Symbols and Abbreviations

AEA- Europe: Association for Educational Assessment- Europe

ALES: Academic Personnel and Graduate Education Examination

ALTE: Association of Language Testers in Europe

CCA: Constant-Comparison Analysis

CCC: Council for Cultural Cooperation

CEFR: Common European Framework of Reference for Languages: Learning, Teaching, Assessment

CILA: Commission Interuniversitaire de Linguistique Appliquée

CoE: Council of Europe

CoHE: Council of Higher Education

DGS: Vertical Transfer Examination

DIALANG: An Online System of Diagnostic Language Assessment

EALTA: European Association for Language Testing and Assessment

EAQUALS: European Association for Quality Language Services

EAP: English for Academic Purposes

ECC: European Cultural Convention

EFL: English as Foreign Language

ELL: English Language Learner

ELP: European Language Portfolio

ELT: English Language Teaching

ENLTA: European Network for Language Testing and Assessment

ESL: English as a Second Language

EU: European Union

Europass: European Skills Passport

FE: Formal Education

GEPT: General English Proficiency Test

IELTS™: The International English Language Testing System

ILTA: International Language Testing Association

IRT: Item Response Theory

KPDS: Foreign Language Examination for Civil Servants

KPSS: Public Personnel Selection Examination

LOA: Learning Oriented Assessment

MEBBIS: The Ministry of National Education Data Processing Systems

MoNE: Ministry of National Education

NFE: Non-Formal Education

ÖSYM: Measuring, Selection and Placement Center of Turkey

ÖZ-KUR-DER: Association of Private Educational Institutions and Study Centers

PTE Academic™: Pearson Test of English Academic

QA: Quality Assurance

RLDs: Reference Level Descriptors

SEM: Standard Error of Measurement

TESOL: Teaching English to the Speakers of Other Languages

TOEFL®: Test of English as a Foreign Language

TOEIC®: Test of English for International Communication

TUS: The Turkish Medical Specialty Examination

UN: United Nations

UNESCO: United Nations Educational, Scientific and Cultural Organization

ÜDS: Inter-University Foreign Language Examination

YDS: Foreign Language Examination

YÖKDİL: Higher Education Institutions Foreign Language Examination

YÖS: The Examination for Foreign Students

Chapter 1

Introduction

Background of the Study: A Personal Perspective

As a former English language teacher at a public school, and currently a Research Assistant in the Department of English Language Teaching (hereafter ELT) at a state university, I am quite familiar with the characteristics of some fundamental English language teaching and testing practices in Turkey. Based on my personal experiences and observations, I can safely state that grammar-oriented testing (Canale & Swain, 1980; Chastain, 1988; Lightbown & Spada, 1990; Morrow, 2012; Richard-Amato, 1988) and structural rules-based assessment formats (Chamot & O'Malley, 1987; Ellis, 1993; Skehan, 1996; Swain & Lapkin, 1995) were once in common use in most of the educational settings across the country. In some educational environments, they are still in use as those implementations seem more practical, or due to teachers' attitudes of resistance to change in their culture of teaching. However, today, European countries refer to the principles and guidelines presented in the Common European Framework of Reference for Languages (henceforth CEFR) (Council of Europe, 2001) as the base in language teaching, learning and assessment, which blossoms interest also in Turkey. In fact, at the 20th session of the Standing Conference of the Ministers of Education of the Council of Europe (henceforth CoE) in Cracow, Poland, it was decided to use the CEFR descriptors commonly as well as to disseminate the use of the European Language Portfolio (henceforth ELP) as a self-assessment tool across Europe (CoE, 2000).

As one of the member states to the CoE, the Turkish Ministry of National Education (henceforth MoNE) has been using the Common Reference Levels defined by the CEFR in order to underpin all the teaching, learning, assessment and certification credentials (Mirici, 2015). Hence, the picture of implementation favors the use of communicatively-oriented curriculum grounded upon an action-oriented approach in foreign language teaching, and relevant testing and assessment practices. Taken together, these steer my interest into this research as the use of the European standards are the basic premises for good practice in foreign language education, specifically, in testing and assessment.

Motivation for the Study

I have been trained as a teacher of EFL in higher education institutions in Turkey. At that case, the reference documents of the Threshold Level (van Ek & Trim, 1990) and the tenets of communicative language teaching have been discussed and highly recommended. Therefore, I am well cognizant of the fact that the European standards are taken as the canons for good practices in teaching, and testing English as a foreign language. Enabling goodness in practice, the proficiency benchmarking in English has also been revised, and added global scales of the CEFR to create a balance between the content and performance standards (Little, 2007). Within years, it has turned out to be more practical as to the previous case; however, this time, heavy reliance on tests has popped up as a burden for students as the judgment is made according to the results they have gained through assessment procedures. Correlatively, the assessment conducted is bounded to quantity rather than quality. Therefore, what learners can do with the functional skills necessitated by the task seems more important than how well learners perform in the sense that they can effectively and efficiently use what they acquire as language skills (De Jong, 2004; Hulstijn, 2007). That is why after those years spent on English language (now it is 11 years until undergraduate education from 2nd to 12th grade), and hours of study ranging between 2 and 4 per week, many are still unable to have a simple act even in daily life conversations. Assuming that they have had adequate grammatical and lexical knowledge, it is to be as easy as pie for students to have a good command of language. But this is not the case. Thus, many of the learners have decided to take further English language education by means of language schools/ courses, study centers and/or other private institutions.

Substantially, this is the point where this dissertation sprouts up. The situation I have depicted above is not solely common in Turkey as there is an ongoing increase in the demand of English language learning through private institutions. According to the British Council's report, it is supposed that by the year 2020, the number of the adult English Language Learners (henceforth ELLs) is expected to rise to about 2 billion from 1.5, meaning that 1 out of 4 is to be using the language across the world (Pearson English, 2014). For Graddol (2006) nearly a third of world population are expected to learn English simultaneously. This expectation of significant growth is the case for English both within and outside English-speaking

countries. For this reason, Turkey as a non-English speaking country and with its EFL context, holds English as a part of school curriculum, and supply courses paid for privately in language learning centers.

In essence, I am motivated to scrutinize whether the CEFR might replenish a fundamental basis for the reconsideration of testing and assessment practices in terms of teaching and learning English in Turkey. Of particular interest, the notion of progression in testing and assessment by some European guidelines for non-formal educational settings is at the core as they are the centers enclosing a great number of English language learners for many reasons. As the ratio of auditing of non-formal educational settings is rather low as to that of formal education, how the Framework is received by non-formal English language schools within the concept of testing and assessment is of utmost importance regarding these institutions as the arteries of Turkish EFL context and education economy.

Research Objectives

The overall aim of this research is to probe into the testing and assessment practices of English language schools in Turkey, which are listed under the heading of non-formal educational institutions. Supposed to do so, how well they trace the applications and basic principles designated by the CEFR, the criteria defined by the European Association for Language Testing and Assessment (hereafter EALTA), the guideline assigned by the International Language Testing Association (hereafter ILTA) and the standards set by the Association of Language Testers in Europe (hereafter ALTE) will be explored. Starting from the very beginning with the decision-makers at these private institutions, this study aims to define the testing and assessment practices of non-formal English language schools within the boundaries of predetermined sub-criteria which will further be touched upon in a more detailed way. Additionally, the importance of assessment in education and the application of European standards in educational assessment with its broadest sense will also be highlighted with the help of the European Framework of Standards for Educational Assessment by the Association of Educational Assessment- Europe (hereafter AEA-Europe).

Research Questions

Laying the emphasis on the research reported in this dissertation, the perceived gap in the literature is postulated to be filled with the answers to the research questions that come into picture as below:

1. Do the testing and assessment practices of non-formal English language schools in Turkey comply with the criteria designated by the EALTA?
2. Do the testing and assessment practices of non-formal English language schools in Turkey correspond to the standards set by the ALTE?
3. Do the testing and assessment practices of non-formal English language schools in Turkey fit the guidelines assigned by the ILTA?
4. What is the role of testing and assessment in Turkey's system of education in the light of the standards set by AEA-Europe?
5. What is the general paradigm of a sample of leading professionals from selected non-formal English language schools in Turkey (i.e. decision-makers, testing office, English language teachers) on the implementation of testing and assessment procedures as defined by the European guidelines?
 - a. Do the testing and assessment practices of selected non-formal English language schools in Turkey differ from each other within the scope of pre-determined European guidelines?
 - b. What are the viewpoints of the directors from the selected private institutions and ÖZ-KUR-DER on the utilization of the European guidelines in testing and assessment practices?

Significance of the Study

Current researches in the field are closely in touch with the implementation of the Framework in language testing and assessment in formal educational settings (Alderson & Huhta, 2005; Cumming, 2009; Davidson & Fulcher, 2007; Green, 2017; Hasselgreen, 2005; Ilc & Stopar, 2015; Little, 2005; Martyniuk, 2010; Stoyhoff, 2012; Tannenbaum & Wylie, 2008; Taylor & Geranpayeh, 2011) Stated as a limitation, examining a wider range of curricula under the influence of the CEFR, namely not school-based and non-formal educational environments, would surely broaden the viewpoints. That is the point that magnifies the significance of this study. Molded

with semi-structured interviews with decision-makers, together with the reports gathered from the English language teachers, test designers and/or examination providers, this study probes into the utilization of the European standards in testing and assessment implementations in use.

On the other hand, English language teachers rendering service to non-formal education (hereafter NFE) platforms are generally busy, and are not assumed to reply in the affirmative to every research-related invitation they receive. Particularly, if there is not any benefit resulted directly from the research, the language teachers prefer not to be involved in any study for this good reason although a kind of privacy legislation is put in place. Herein, it is a crystal-clear fact that this study is to lend assistance to those who are responsible for testing and assessment procedures, and thereafter, to shed light on related testing and assessment practices within the context of some European standards.

Assumptions

It is assumed that this study will contribute to foreign language research within the scope of testing and assessment. The framework, guiding principles and standards are used as a basis for the appropriateness of testing and assessment practices to those European standards. Contrary to the ordinary, what is more to the point is that the testing and assessment practices of the private institutions rendering non-formal English language education are taken as the core instructional context for this study. Accordingly, the role of setting standards in testing and assessment, and the use of standardized tests are assumed to blossom as the needs for subsequent practices.

Limitations

Only 3 private institutions as non-formal English language schools are included in this study. Although they are the most commonly preferred, widely-known and influential language learning centers of private education sector in Turkey, the number could be increased. Additionally, the students are not included in the study. Their understanding of testing and assessment practices could be included; however, it is hard to control such a wide range of variables all at once. Besides, this study embarks on the language testing and assessment practices of non-formal English language schools in Turkey; however, the test formats in use

are not analyzed. This is due to the fact that analyzing test formats in use could make the scope of this dissertation derailed as such kind of analysis requires the implementation of some specific matrices in order to ensure the essentials of testing and assessment, such as validity, reliability, practicality, and the like.

Definitions of the Key Terms

English language schools: The private institutions, centers and/or courses providing English language education.

International Language Testing Association (ILTA): It is the Association that aims “to promote the improvement of language testing throughout the world” (ILTA, 2008, p. 1).

Non-formal education (NFE): “Any form of systematic learning conducted outside of a formal organization” (Jarvis, 1987, p. 21).

The Association of Educational Assessment- Europe (AEA- Europe): It is the Association that aims “to act as a European platform for discussion of developments in educational assessment, fostering co-operation and facilitating liaison between organizations and persons active in educational assessment across the whole of Europe” (AEA- Europe, 2013, p. 2).

The Association for Language Testers in Europe (ALTE): It is the Association that aims “to promote the transnational recognition of language certification, and to establish and maintain common standards for all stages of the language testing process in Europe” (ALTE, 2012, p. 1).

The Common European Framework of Reference for Languages (CEFR): “The CEFR provides a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe” (CoE, 2001, p. 1).

The European Association for Language Testing and Assessment (EALTA): It is the Association that aims “to promote the understanding of theoretical principles of language testing and assessment, and the improvement and sharing of testing and assessment practices throughout Europe” (EALTA, 2006, p. 1).

The European Language Portfolio (ELP): “The European Language Portfolio (ELP) provides a format in which language learning and intercultural experiences of the most diverse kinds can be recorded and formally recognized” (CoE, 2001, p. 5).

The Organization of the Dissertation

The organization of this dissertation is wheeled around two major scopes. Accordingly, the European standards in language testing and assessment are touched upon as the first main consideration. Within, the European standards are framed by the guidelines of the CEFR, EALTA, ALTE, ILTA and AEA- Europe respectively. Each of the European standards aforementioned is probed in detail with their origins, development, main objectives, specifications and core elements. Besides, a review of recent studies conducted on the utilization of each of these European standards are elaborated.

Secondarily, the national standards in language testing and assessment regarding the case in Turkey are reviewed. Herein, the Turkish national education system is briefly mentioned, and laced with the language testing and assessment practices in formal and non-formal educational settings in tow. For the formal educational settings, ÖSYM, as the Measuring, Selection and Placement Center conducting large-scale language examinations in Turkey, is highlighted. Within the scope of non-formal educational settings, ÖZ-KUR-DER, as the Association of Private Educational Institutions and Study Centers in Turkey is elaborated in detail. To note more, other relevant units of testing and assessment are briefly discussed, as well.

Following these, methodology part is composed of the information on the themes of research design, population as the sample size and settings, materials used as data collection instruments, data analysis procedure including statistical techniques laced with demographic information, descriptives and related testing of assumptions, and ethical considerations respectively.

Moreover, findings and results are elaborated within tables and estimations gathered through statistical analyses, paving the way towards conclusion and discussion part. Herein, the results are discussed, and an overview of the study is embellished with the studies placed in the literature. Around three major scopes, the utilization of the European guidelines in testing and assessment practices by

selected private institutions are emphasized together with some pedagogical implications and recommendations for further research, which is ended with a detailed conclusion section.



Chapter 2

The European Standards of Language Testing and Assessment

Introduction

In this chapter, the author of this dissertation discusses some European standards for establishing a common practice in language testing and assessment. Accordingly, this part is composed of five sub-headings, namely the Common European Framework of Reference for Languages, the European Association for Language Testing and Assessment, the Association for Language Testers in Europe, International Language Testing Association, and the Association of Educational Assessment- Europe. Each of the European standards listed is probed in detail to draw a general picture of language testing and assessment across Europe.

The CEFR: Origins, Content and Development

In this very first part, the origins, content and development of the CEFR are touched upon. Accordingly, the development of the Framework is briefly explained. Then, a brief outline of the European content, language policies and their contribution to the creation of the CEFR by means of reference level descriptors and certification systems is drawn. Following that, the ELP as a self-assessment tool, and the Manuals as a reference supplement for language testing and assessment practices are mentioned respectively. In the last section, a review of recent studies conducted on the utilization of the Framework in language testing and assessment is introduced.

The development of the framework. The concept of the CEFR dates back to the 1970s. However, it was officially launched in 2001. Within a historicist point of view, Europe inherited a wreck after the Second World War. Not only economy, but also international relations were in ruins. Accompanied by the Cold War afterwards, European nations were not able to have a contact with each other. The situation is best summarized by the words of Trim (2005) as “under such conditions, language teachers became quite out of touch with the up-to-date realities of the languages and cultures they were teaching and concentrated their attention on puristic formal correctness and the heritage of national literature” (p. 13).

Such tragic events and post-war clouds on Europe changed the Europeans' views of thinking. Accordingly, Valax (2011) states that within the scope of competition amidst the United States, Japan and other emerging powers like China, India and Brazil followed by the harsh period of renewal, Europe was nourished by the post-war Europeans' beliefs to unite against the reiteration of a blue funk of the war. This was because gaining a robust entity was believed to fasten the ties among European nations and toughen Europe's stance against the forthcoming challenges of globalization. Notwithstanding, in *pari passu* significance, the Europeans' need for unity was to be molded by a number of characteristics, values of a variety of perceptions, and language and cultural diversity laced with mutual understanding and cross-tolerance. Accordingly, the context of post-war Europe, and the seek for unity and cooperation among European nations led to the establishment of a variety of organizations such as the CoE, and European Cultural Convention (henceforth ECC) in order to appreciate the pros of getting together under a single but a much stronger entity, which later paved the way for the creation of the Framework.

The CoE was founded in May 1949 with the core values gathered around human rights, democracy and the rule of law. Covering millions of citizens and 47 member states, the CoE acts as an advisory body for European countries in order to foster cooperation, awareness of respect and unity. It also reverbs in the Framework as the CEFR:

“serves the overall aim of the CoE as defined in Recommendations R (82) 18 and R (98) 6 of the Committee of Ministers: ‘to achieve greater unity among its members’, and to pursue this aim ‘by the adoption of common action in the cultural field’” (CoE, 2001, p. 2).

Seeing that the Council lends wings to the quality improvement of communication among European nations of different linguistic and cultural backgrounds, the objectives of the Council are pursued in case of the maintenance of a strong unity between its members. Therefore, European nations were triggered to determine policies not only within the field of modern language learning and teaching, but also those regarding cultural development. Considering the aims of the CoE, the methods are to be executed in order to accomplish those aims in the context of human rights and fundamental freedoms. In this sense, the most significant achievement of the CoE was the adoption of the European Convention on Human Rights in 1950.

Accordingly, it was announced in the inaugural of the Convention that fundamental freedoms as the basic requirements for justice and peace in the world could only be maintained through effective steps taken towards mutual understanding, tolerance and democracy. Reaffirming this belief, the likeminded governments of European countries that once shared much in common were fostered to use collective enforcement to the path towards richness in language and cultural diversity, and unity in cooperation. It is also issued by the ECC, pursuant thereto the context with primary focus on unity and diversity asserted by the CoE. Ratified by all 47 member states of the CoE and by Belarus, Kazakhstan and the Holy See, its signature is accepted as the symbolic key for the Bologna Process, which is constituted by subsequent ministerial agreements and meetings amidst European countries within the framework of comparability of the standards. In this sense, the ECC is the terminus a quo for the work of the CoE in the field of modern languages, and for cultural co-operation throughout Europe. Concomitantly, in December 1961, the Council for Cultural Cooperation (hereafter CCC) was brought into existence by the Committee of Ministers of the CoE, embodying four committees which are incumbent upon culture, cultural heritage, education and higher education.

In the 1970s, the need for a new methodology and approaches for the definition of the objectives and content (Saville, 2005) mushroomed the birth of a new approach as an offspring against audio-lingual and visual methods. Inspired by the Symposium on Languages in adult education held at Rüsclikon in 1971, a research group made up of John Trim, Jan van Ek, David Wilkins and René Richterich was ensued by the Committee for Out-of-School Education in order to check the applicability of a unit/credit system for adult language learning. Within this system, the subjects were to be taught in parts, albeit not globally. Accordingly, John Trim rolled up the sleeves to prepare a list of function types and speech acts. Herewith, David Wilkins published these in 1973. Within the same year, an article entitled 'The Linguistic and Situational Content of the Common Core in a Unit/Credit System' (Wilkins, 1973) was published in which the background of the system undertaken for adult language learning was explored.

By 1977, all levels of the educational system were significantly influenced by the string of ideas evolving around the functional-notional approach and

construction of foreign language courses as to the characteristics of the learners. After the presentation of the Threshold Level, laced by the unit/credit system in adult language learning, it was decided that the research group was taking promising actions towards foreign language education; therefore, a new project was to be set. It was Project No. 4: 'Modern Languages, improving and intensifying language learning as a factor making for European understanding, cooperation and mobility', which took time between the years of 1977 and 1981. In this project, the ultimate aim was to consider the learners' needs and characteristics to improve the ability to use the language and control his/her own progress, and to make it probable and accessible for all sections of the population to enhance understanding, cooperation and mobility throughout Europe (Girard & Trim, 1988). This time, "the principles developed by the unit-credit group were applied in projects across the different sectors of general secondary, vocational and adult education, as well as in migrant education" (Trim, 2001, p. 4).

However, the birth of the Framework in company with that of the European Language Portfolio (ELP) was accepted at the Rüsclikon Symposium. Initiated by the Swiss federal government and respective organizations, an Intergovernmental Symposium under the head of 'Transparency and Coherence in Language Learning in Europe: Objectives, Evaluation and Certification' was held in Rüsclikon in 1991. The main objective of the symposium was to relate language programs and examinations through the agency of a common framework of reference (North, 2005). Thus, language programs with language examinations in tow, would merge under a common mental framework to attain the main themes of the symposium: 'transparency and coherence'. In fact, the idea of having a common system in language education was formerly revealed as Trim already "put forward the draft of a system in 1977 and ... tried to get a unit developed to establish and administer it" (Saville, 2005, p. 278); however, there was a strong inquietude of European centralism, especially in Scandinavia. Thanks to the efforts of Switzerland, the notion came to the fore again as Switzerland stated that "the degree of educational and vocational mobility means that people are always having to evaluate qualifications which they don't know anything about" (Saville, 2005, p. 279). In this sense, between the years of 1989 and 1990, a group of emissaries from Eurocenters and a study group from the CILA (Commission Interuniversitaire de

Linguistique Appliquée) gathered to localize the linguistic competences alleged by different forms of certification systems and examinations so that they could examine the probability of setting a transparent and a common system and/or a model for exams, diplomas other certifications.

In the light of these, the objectives determined by the Framework were (Trim, 2005):

- *“to promote and facilitate co-operation among educational institutions in different countries;*
- *to provide a sound basis for the mutual recognition of language qualifications;*
- *to assist learners, teachers, course designers, examining bodies and educational administrators to situate and co-ordinate their efforts” (p. 14).*

After series of revisions and amendments, the final version was announced at the ‘European Year of Languages’ organized jointly by the CoE and European Union (henceforth EU). This final version was published both in French and English as the Framework, and presented with the ELP in 2001 together with the guides and manuals developed for the Framework.

The reference level descriptions and certification systems. The Framework proposes linguistic descriptors molded with acquired (sub) competences to define a trajectory for language learning. These descriptors are not language-specific, albeit applicable to all across-to-board implementations. Accordingly, the descriptors grade the booming skill-mastery by means of a six-level scale (A1, A2, B1, B2, C1 and C2). Nevertheless, for the practitioners such as teachers, course material designers and textbook writers, the levels of specifications of the CEFR may seem to be highly cosmical. For this purpose, the CEFR specifications have been examined one by one for each language. As a result, reference level descriptions generated brand-new are grounded upon the linguistic forms, mastery of communication, socio-linguistic competence and other competences described by the CEFR. Leading to the development of the Reference Level Descriptions (hereafter RLDs) for national and regional languages, this conveyance of the CEFR into a chosen language has blossomed as an outline of the common general principles developed “in order to give these reference level descriptions for individual languages a degree of scientific status, and a social audience compatible with their aim” (CoE, 2005a, p. 6).

Taken as the milestones for the development of national and regional language programs at a common core, the RLDs could be used for different languages in order to share common tools; therefore, the language teaching programs could be in association with each other. To add more, the descriptions are for all European languages; albeit not available to solely one specific language, specifying the notion that no language is superior to another. Enabling the language knowledge accessible to all competence types at any level, these descriptors directs the language teaching and learning in a more transparent way; on top of it all, the RLDs are certified by the reference instruments, as well.

Enshrining a transparent and novel way for language learning, the CEFR has also led to improvements in the field of assessment by labelling the proficiency levels in a more specified way, compared to the traditional practices which were once prevalently in use. The levels in the CEFR are far from just having numerical data, at least more meaningful than it. Trim (2005) states that:

“The scales and descriptors have been of special interest to authorities who want to situate their language qualifications relative to those of others, and to the ‘users’ of qualifications gained in other systems, such as employers in deciding who to appoint to jobs involving language use to a greater or lesser extent and educational authorities in establishing entry requirements for courses at different levels” (p. 17).

Within the field of assessment, the ELP is the first as a self-assessment tool with the intention of providing learners assistance to better understand their progress. It also promotes international mobility by facilitating the understanding of the learning process. Parallel to the development of the CEFR and ELP, the other certification documents are guides and manuals which are to show the implementation of the CEFR. To elaborate, the 1996 version of the CEFR, which was accompanied by the eleven guides, was modified by ‘A Guide for Users’. Later, the final version was announced in 2009 as ‘A Manual for Relating Language Examinations to the CEFR’ (CoE, 2009a). Backed up with a series of reference materials such as videos, DVDs and/or CDs, the ‘Reference Supplement’ (CoE, 2009b) is comprised of multifunctional information on sample calibrated performances in order to nudge relevant persons who are responsible for examination in direction to make better judgment.

However, Figueiras (2007) proposes that “[as] early as 2001, de Jong’s unpublished presentation at the Barcelona Conference of the ALTE Conference [sic]

listed the dangers of rash and unreliable claims of linkage of examination levels to the CEFR levels” (p. 673). On the other hand, until 2007, there was the appearance of “countless bodies purporting to deliver certificates or diplomas based on the CEFR levels or to guarantee that such an examination or qualification demonstrates linguistic competence at a specific CEFR level” (Bonnet, 2007, p. 671). In the 2005 survey (CoE, 2005b), the CEFR was approved as being useful “in the domains of testing / assessment/ certification (2.70 on a 0-3 scale)” (p. 3-4). As a result of the 2006 survey (Martyniuk & Noijons, 2007), the CEFR was approved as being useful in the context of curriculum development at the ration of 87% (26 out of 30 representatives). Although this is the case, there is a systemic risk of using the CEFR for assessment without calibration. Even so, the touchstone in using the CEFR as a guide for teaching, learning and assessment practices is that “there is not and never will be an authorized interpretation of the CEFR. That openness is the secret of its success” (North, 2014, p. 5). In this sense, North, Martyniuk & Panthier (2010) assert that:

“The CEFR is a concertina-like reference tool that ... educational professionals can merge or sub-divide, elaborate or summarize, adopt or adapt according to the needs of their context... It is for users to choose activities, competences and proficiency stepping-stones that are appropriate to their local context...” (p. 4).

Concerning these, there has been a rapid change towards the alignment of qualifications as to the standards set by the CEFR. This is followed by the process through which examinations have been related to the CEFR as described in the Manual. By reporting the outcomes of the learning process into a symbolic format by means of levels on the scale, any educational system may be controlled somehow. Because the results are interpretable within the terms of levels proposed by the scale itself. This is why any ‘CEFR-aligned’ document, either a test or an exam, is preferred on the grounds that it is to be good.

To add more, although there is the presence of false interpretations and a great number of problems in terms of linking and/or aligning examinations to the CEFR, there is the absence of a committee composed of experts to validate and deal with the certification problems. On this point, Alderson (2007) states:

“The Council of Europe set up a so-called Validation Committee to vet (or rubber-stamp) the large number of European Language Portfolios (ELPs) that were developed in the late 1990s and early 21st century. Unfortunately, despite the greater influence of examinations on the curriculum—and on

lives—the Council of Europe has refused to set up an equivalent mechanism to validate or even inspect the claims made by examination providers or textbook developers” (p. 661).

Alderson (2007) also notes that the European Association of Language Testing and Assessment (EALTA), to the best of its independence, is the only organization that is responsible for dealing with these types of problems, albeit leastwise.

The European language portfolio (ELP). As briefly touched upon above, the ELP is a tool for learners providing self-assessment, so that learners are able to keep track of their results accomplished, qualifications gained and competences acquired. This record embodies all learning activities at any level, either at or outside the school. It is why learning is described within the framework of languages and intercultural experiences as taking place all life-long. In 2001, after an attempt to ensure accreditation, the ELP was officially introduced in tandem with the CEFR. In an attempt to “document their progress towards pluri-lingual competence by recording learning experiences of all kinds over a wide range of languages” (CoE, 2001, p. 20), the ELP aims to provide support for learners in the field of language studies.

Within the scope of the ELP, there are three parts, namely the ‘Language Biography’, ‘Language Passport’ and ‘Dossier’. Briefly, the Language Biography keeps the records of learners’ language learning and intercultural experiences in both formal and informal educational environments including school context, experiences gained through exchange programs and working area. On the other hand, the Language Passport points individual’s proficiency in languages at a certain period of time. It includes the summary of one’s competences of languages learnt, describing the overview by means of skills and the common reference levels. One has the opportunity to update and record regularly his/her Passport with the help of qualifications and diplomas received, self-assessment reports filled up and all kinds of intercultural experiences gained. In this vein, it is intended to support individuals to opt in the learning process from planning phase to the stage of evaluation. The inclusion of learners’ skills, experiences and achievements within the scope of foreign languages is reflected by their selection on the Dossier. The Dossier is the storage box for learners’ progresses and updates during the time of individual growth (CoE, 2006; Rehorick & Lafargue, 2005).

In the light of these, there are some critical aspects regarding the Portfolio. The Portfolio has the authority to reflect the learning processes of different languages at the same time, contrary to what is believed just as the recording of qualifications obtained in formal educational contexts. To add more, the Portfolio can be shaped according to age and/or local contexts on condition that the CoE's Validation Committee does approve the same standards. Therefore, coherence amidst different groups is ensured in order to be gathered under a common core proposed by the CEFR. To set an example, the CoE's Validation Committee has approved one hundred and thirteen models of Portfolios as valid by year 2010 (Valax, 2011). Nevertheless, the validation process has ended with the discharge of the Validation Committee in the same year. Instead, an on-line registration system approved as a part of self-declaration by the CoE's Language Policy Division has been brought into use since April 2011. Besides, a wide variety of documents have been developed to assist portfolio developers, teachers and teacher trainers. To add more, the electronic form of Language Passport for adults, namely the European Skills Passport (henceforth Europass), was put forward by the CoE and EU in 2004. It can either be completed on-line or downloaded from the website. By the way, the first electronic ELP has been developed by the European Association for Quality Language Service (henceforth EAQUALS) and ALTE, and herewith accredited.

The manuals. As noted by Coste (2007), one of the authors of the CEFR, the Framework has a notable influence on language assessment; henceforth; the alignment of the language tests to the CEFR has drawn more attention than the other aspects of the Framework. In this context, a bunch of tools are introduced to assessment providers and/or practitioners who are interested in language testing and assessment. Amidst them, first one is the 'Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment' (CoE, 2009a). Besides, a technical 'Reference Supplement to the Manual for Relating Examinations to the CEFR' has also been introduced (Banerjee, 2004; Eckes, 2009; Kaftandijeva, 2004; Verhelst, 2004a, 2004b, 2004c, 2004d) to enable standardization in developing tests, and aligning them to the Framework. Alongside, a number of materials exemplifying the levels of the Framework, and content analysis grids for each of the language skills have been developed. Additionally, some attempts to develop RLDs for English and some other

languages have also been introduced for the benefit of language testers, test developers and practitioners in the field.

Additionally, on behalf of the Language Policy Division of the CoE, the ALTE has introduced the 'Manual for Language Test Development and Examining' to be used with the CEFR effectively within their own contexts, and by their own objectives (CoE, 2011). Herein, it is to be noted that the 'Manual for Relating Language Examinations to the CEFR' together with its 'Reference Supplement have been designed to address a general approach for the alignment of the tests to the CEFR in order to set standards and a number of options. On the other hand, the 'Manual for Language Test Development and Examining' has been designed as a complementary document for the Manual previously mentioned. It centers on some other aspects, which are not touched upon in the other Manual; therefore, it is accepted as the revised version of the document of 'Users' Guide for Examiners' (CoE, 1996).

Within the scope of this Manual, fundamental considerations such as language proficiency, validity, reliability, ethics and fairness are highlighted. Following these, the process of test development together with test requirements, test specifications and some practical considerations are presented in tow. Correlatively, the process of assembling and delivering tests is identified. Within the scope of assembling tests, producing and managing materials, commissioning, item writer recruitment and training are probed in detail by elucidating the assessing requirements. Besides, quality control analyses are emphasized within in order to construct tests in tune with the Framework. Moreover, within the scope of delivering tests, the process is scanned by means of sending and returning materials together with administering the tests, which are followed by the steps of marking, grading and reporting of the results respectively. The Manual is also appertaining to the monitoring process and test review.

A review of recent studies on the utilization of the framework in language testing and assessment. With a review of studies conducted on language testing and assessment within the scope of the CEFR, it could be stipulated that formal educational settings have generally been at the core. Accordingly, skills-based language assessment, self-assessment and aligning tests to the CEFR have long been a common focus of the studies conducted. On the other

hand, criticisms on the Framework are also put into use by some studies conducted, though. Below, a compile of aforementioned studies has been introduced to set a panorama to the review of literature.

To begin with, Little (2005) addresses the Framework and ELP, and reports on a project developed for defining an English as a Second Language (ESL) curriculum by involving non-English speaking students of Irish primary schools within. Herein, he has developed a version of the ELP in order to gather the learners' judgments in the assessment process, and has reported that self-assessment has a key role in learner-centered approaches to language teaching. In a similar vein, Hasselgreen (2005) has focused on the assessment of young learners by the utilization of the CEFR and ELP. With a special concern on the assessment practices, it is probed to what extent the special needs of the young learners are catered for in some European schools. As a result, it is reported that the adoption of the CEFR and ELP has been directing the language assessment practices within the scope of young language learners in Norway, and leading the path towards embracing the Framework as a more common way of young language learners' assessment in present day Europe. For the Taiwanese educational context, Wu (2008) has asserted that with the adoption of the CEFR in 2005, and the launch of the General English Proficiency Test (GEPT) in 2000, there has been a critical change in Taiwan. Herein, she has reported that some Taiwanese students and teachers as the participants of her study have a positive tendency to use the Framework for language testing and assessment. As an implication for language teaching and assessment in Korea, Finch (2009) has suggested the adoption of the CEFR as well as the Europass as an offspring of the 'Europass Project' (European Communities, 2009) in order to develop a 'Koreapass' or an 'Asiapass'.

For setting standards to relate language examinations to the CEFR, Tannenbaum and Wylie (2008) have worked on linking English language test scores with the Framework by using the scores of three tests, namely Test of English as a Foreign Language (TOEFL®) iBT, Test of English for International Communication (TOEIC®) and TOEIC *Bridge*™ tests. They have linked TOEFL® iBT at the proficiency levels of B1, B2 and C1; TOEIC® at the proficiency levels between A1-C1; and TOEIC *Bridge*™ at the three targeted proficiency levels of the Framework. In a similar context, the Foreign Language Proficiency Test Administered in Turkey

(YDS), named previously as Foreign Language Proficiency Exam for the State Employees at that time (KPDS), has also been analyzed in terms of its appropriateness to the CEFR, taking the years between 1990-2013 into concern (Demir & Genç, 2016). Herein, German language exams are taken to the fore, and it is suggested that the equivalences presented by the tables of the Council of Higher Education in Turkey (henceforth CoHE) regarding A1 and A2 proficiency levels are to be removed.

As the practices in language testing and assessment have been changed with the arrival of the Framework, Inbar-Lourie (2008) has focused on language assessment courses in order to build a basis for language assessment knowledge. She has proposed language assessment courses to establish a core knowledge in language assessment, and to meet the demand for language assessment literacy. To raise the university level students' awareness of their own speaking skills, Glover (2011) has suggested the use of CEFR level descriptors. More recently, Mirici and Kavaklı (2017) have analyzed the courses of an M.A. program of an ELT department in Turkey in order to promote a deeper look at teaching CEFR oriented practices effectively for students to internalize the Framework better. Not to mention, Malone (2017) has proposed professional training in order to develop a better understanding of language assessment taking the Framework at the core of language testing and assessment practices. Similarly, Kavaklı (2017a) suggests the adoption of the Framework to develop EFL teacher candidates' assessment literacy, as well.

To note more, Alderson and Huhta (2005) have purported a suite of computer-based diagnostic tests based on the Framework. Herein, they have suggested DIALANG, as an on-line language assessment system, which has also a basis on the CEFR with amount of 14 tests available in European languages. The difference, here, is that DIALANG is used to diagnose language skills by providing feedback to users instead of pointing out their proficiency levels. Thus, it is neither an examination nor a certificate issuer, but offers validated language tests on a skills-based format. In a similar way, a CEFR-based testing system for Chinese language proficiency has also been developed as a computerized adaptive testing system (Wang, Bor-Chen, Tsai & Liao, 2012). In a very recent study, some tasks have been explored to develop classroom language assessment benchmarks for Japanese EFL teachers (Kimura, Nakata, Ikeno, Naganuma & Andrews, 2017).

Herein, the aforementioned tasks have been framed by the guidelines of the Framework, as well.

In addition to these, the Framework has been centralized for skills-based language assessment. Herein, it is probed whether the reading and listening tests are aligned with the CEFR by dint of the results gathered by the experience of the 'Dutch CEFR Construct Project' (Alderson et al., 2006). The outcomes have been reported as the promising nature of the Dutch CEFR Grid developed, albeit mushrooming the need of improvement on test task levels and test specifications. Taking the Framework as the basis, second language vocabulary assessment is probed by Read (2007) including the 'Academic Word List' and the vocabulary list of the 'British National Corpus'. Similarly, linguistic competences of Dutch as a second language learners with B1 and B2 levels of speaking proficiency have been analyzed by using the Framework (Hulstijn, Schoonen, de Jong, Steinel & Florijn, 2011). In doing this, they have underpinned the CEFR by applying 'Overall Oral Production Scale' of the Framework (CoE, 2001). As a result, they have concluded that the differences in B1 and B2 learners' lexical and grammatical knowledge are most likely to be a matter of degree, albeit not that of domain. For the testing of higher values, the Framework is applied by Taylor and Geranpayeh (2011) in order to assess listening for academic purposes. They have tried to operationalize the test construct of L2 academic listening ability, and have recommended the Framework as a guide for test designers. For assessing speaking skills, Roca-Varela and Palacios (2013) have pinpointed the general guidelines of the CEFR. Accordingly, for the reformulation of the assessment of the oral skills, the nature of the Framework encompasses different types of tasks and marking systems. Very recently, for the assessment of meaning, Purpura (2017) has investigated how L2 testers conceptualize meaning with regard to L2 proficiency through the use of the 'can-do' statements. Herein, it is highlighted how the expression and comprehension of meaning are internalized in L2 assessment formats. Hence, the meanings which the testers want to test, and those which are implicitly assessed have been blossomed consequently.

Contrary to those listed, some studies have been conducted to mark the demonstrable weaknesses of the CEFR in terms of designing language tests, though. Confirming this, Weir (2005) asserts that the current version of the

Framework is not adequately comprehensive for language testing as the contextual parameters of the 'can-do' statements hamper linking separate assessments, especially through social mediation. Similarly, Alderson (2007) has stated that there is a need for more research in the utilization of the CEFR in language testing and assessment. In the same vein, Hulstijn (2007) has associated this need with a 'shaky ground' of the Framework in terms of qualitative and quantitative language proficiency dimensions. Herein, Davidson and Fulcher (2007) discuss the flexible language of the Framework in order to investigate the pragmatic use of this language to lead language test development by selecting service encounters, and using A1 level of descriptors. As a result, they have reported that there is a need for a revision on some of the test specifications; however, the Framework is the offset of language test development. Besides, Fulcher, Davidson and Kemp (2010) have studied on a scale for the development of speaking tests. They have developed a new scoring instrument for those who are aligning speaking tests to the CEFR, namely 'Performance Decision Tree's. Herein, they have recommended that this instrument would avoid the reification of pre-defined scale descriptors. Similarly, Hulstijn (2011) has marked the failure of the Framework in distinguishing L2 development and L2 proficiency with his study on native and non-native speakers of English, and has come with several implications for a better second language assessment. Recently, Green (2017) has analyzed the process of linking tests of English for Academic Purposes (EAP) to the Framework in a score user's perspective. Accordingly, he addresses specification, empirical validation and standard setting, which are basically the stages suggested by the CoE (2009a). Herein, it is reported by the findings that testing centers do not make much use of the categories of the Framework so as to clarify test content, and the uniformity of the test scores with the CEFR levels.

In reply to above mentioned criticisms of the CEFR, Byram and Parmenter (2012) have purported that the Framework has increasingly been at the core of language learning and assessment practices worldwide. However, its implementation in educational settings is mainly not cognizant of impact studies. Therefore, a very recent study by Piccardo, North and Maldina (2017) reports on 'QualiCEFR', which is a two-year study on integrating both qualitative and quantitative methods together with a Quality Assurance (QA) approach to enhance

the implementations of the CEFR. As a result, promising practices and outcomes for further implementations have been identified as an upscaled version of the Framework. Likewise, in order to bridge the gap between assessment and learning, Alderson, Brunfaut and Harding (2015) have reviewed the implementations of the Framework with a special concern on second and foreign language assessment. One more to note, the Framework has long been criticized with not including the assessment of the pronunciation skills of the language learners. Herein, Isaacs and Trofimovich (2017) have suggested some interdisciplinary perspectives for second language pronunciation assessment by the utilization of the CEFR, as the Framework is currently enshrined in many other educational domains.

Consequently, a panorama to the review of literature in relation with the CEFR, and language testing and assessment is drawn above. Below, the guidelines of the European Association for Language Testing and Assessment are presented together with its mission, principles and guidelines. Last but not least, a review of recent studies on the utilization of the guidelines of the EALTA in language testing and assessment is also framed.

The EALTA: Mission, Principles and Considerations

In this section, the author of this dissertation discusses the mission, principles and main considerations of the EALTA. First of all, a general description of the EALTA is touched on together with its adopted mission. Following that, guidelines for good practice in language testing and assessment in liaison with main considerations within the construction of the EALTA are explained one by one, namely considerations for teacher pre-service and in-service training in testing and assessment, considerations for classroom testing and assessment, and considerations for test development in national or institutional testing units or centers. As the last section, a final note together with the recent studies conducted relatedly is given on the EALTA considerations, taken into account as one of the branches for setting European standards in testing and assessment, which are also adopted within the scope of this study.

The EALTA guidelines for good practice in language testing and assessment. Obtaining participatory status with the CoE in 2008 although founded in 2004, the EALTA acts as a professional association for language testers in

Europe. Besides, the EALTA serves with the financial help from the European Community in order to promote understanding of the theoretical background and related principles in the guise of language testing and assessment. Based on the rationale that Europe is diversified by a bunch of languages, traditions and cultures, such a diversity surely leads to multifariousness in education systems, and so does in traditional way of assessment procedures. In this respect, the EALTA revitalizes testing and assessment practices to be shared and improved within the boundaries of respect in diversity and improvement in quality for the measurement of educational outcomes throughout Europe.

In essence, the need for a European language testing association has popped up with the dissemination of the CEFR and ELP, together with the adoption of language policies projected by the EU and CoE. In this vein, believing the importance of international cooperation for the improvement in the quality of language testing and assessment practices, the EALTA provides individuals, institutions and nations with support to work hand in hand without privilege. By doing this without any diminution of one's cultural identity, the EALTA seeks for independence, internationality, inclusiveness and non-politicalness in practice. Minimizing costs for its members, the EALTA offers membership for all such as teacher educators, students in higher education, teachers, people working at testing units and/or centers, researchers from different field of study and institutions. Besides, the EALTA has organized annual conferences to set an international platform for the sharing of experiences and practices concerning language testing and assessment since 2004. To promote training in language testing and assessment, regional workshops and colloquia, web-based distance courses, special interest groups, reading lists, residential courses and such events are other activities created in the work-stream of the EALTA. Through these activities, it is aimed to increase public understanding, develop links with others who are interested in language testing and assessment, and to engage in activities in order to improve language testing and assessment practices in Europe.

With a view to the 'EALTA Guidelines for Good Practice in Language Testing and Assessment' (EALTA, 2006), they reflect the main objectives of the EALTA addressing three different types of audiences, who will be further mentioned in detail. Adopted in 2006 and translated into 35 different languages, these Guidelines

betoken for those who are involved in (a) the training of teachers in testing and assessment, (b) classroom testing and assessment, and (c) the development of tests in national or institutional testing units or centers. For all aforesaid groups, the general principles assumed to be applied are defined as the respect for the students/examinees, fairness, validity, reliability, responsibility and collaboration among the allies involved. In the light of these, considerations for all groups are scrutinized below on an individual basis.

The considerations for teacher pre-service and in-service training in testing and assessment. In order to create a network for testers in Europe, a project namely 'The European Network for Language Testing and Assessment' (henceforth ENLTA) has been launched. The ENLTA has produced a 'Code of Practice' as a draft form to be revised and reviewed by the working group appointed by the EALTA Executive Committee. Accordingly, with the aim to develop the final version for that Code of Practice, the working group has dealt with creating a set of guidelines in accordance with the EALTA and its mission. Consequently, the EALTA Guidelines have been composed juxtaposing the principles of accountability, inclusiveness and transparency in language testing and assessment. Reflecting these principles, the EALTA addresses three different groups as mentioned above. First one is consisted of the considerations for teacher pre-service and in-service training in testing and assessment. By this way, the EALTA clarifies main considerations for the stakeholders such as trainees, curriculum developers and practicing teachers that are involved in training teachers in language testing and assessment (EALTA, 2006) as follows:

- (1) *"How relevant is the training to the assessment context of the trainees?"*
- (2) *How aware are trainees made of the range of assessment procedures appropriate to their present or future needs?*
- (3) *How clearly are the principles of testing and assessment (e.g. validity, reliability, fairness, washback) related to the trainees' context?*
- (4) *What is the balance between theory and practice in the training?*
- (5) *How far are the trainees involved in developing, trialling and evaluating assessment procedures?*
- (6) *How far are trainees involved in marking or assessing student performances?*
- (7) *What attention is given to the appropriate analysis of assessment results?*
- (8) *What account is taken of trainees' views on the appropriacy and accuracy of assessment procedures?*

(9) How far do assessment procedures used to evaluate the trainees follow the principles they have been taught?" (p. 2).

As seen above, the considerations for teacher pre-service and in-service training for language testing and assessment entail the context of assessment, the awareness of the trainees, the clarity of the basic testing principles, the gap between theory and practice in training, marking and interpretation of the assessment results, the appropriateness and accurateness of the assessment procedure, and the evaluation of the trainees' knowledge on testing and assessment procedure.

The considerations for classroom testing and assessment. Besides the considerations for teacher pre-service and in-service training in testing and assessment, there are also the considerations for classroom testing and assessment. Herein, the EALTA offers clarification for in-class applications within the scope of testing and assessment, taking the stakeholders especially as pupils, and if possible parents. In this respect, the considerations for these aforementioned stakeholders as part of classroom testing and assessment are composed of (1) 'assessment purpose(s) and specification'; (2) 'assessment procedure' and (3) 'consequences' (EALTA, 2006), each of which has its own subcomponents.

To elaborate, assessment purpose(s) and specification embraces the purpose of the assessment, its relation to the curriculum, test specifications, the coverage of the curriculum, how well the assessment purposes are made known, and how well the specifications are discussed. On the other hand, assessment procedures include the design of the procedure, the appropriateness of the assessment procedures to the learners, the ways for gathering information from the learners, the assessment and storage of the learners' information gathered, the efforts for accurate and fair assessment procedure, the promotion of the agreement in marking practices from cross-over applications by other teachers and schools, and learners' views on the assessment procedures. Ultimately, consequences embody the use of the assessment results, the actions to improve learning, the type of the feedback that the students are to get, the ways for learners to make complaints and demand re-assessments, the consequences of the learners' assessment results and those of assessment procedures for classroom practice.

The considerations for test development in national or institutional testing units or centers. Within the boundaries of the considerations for test

development in national or institutional testing units or centers, the EALTA also seeks for answers to the questions listed under the headings of (1) 'test purpose and specification'; (2) 'test design and item writing'; (3) 'quality control and test analyses'; (4) 'test administration'; (5) 'review'; (6) 'washback'; (7) 'linkage to the CEFR (EALTA, 2006). Accordingly, the concerned stakeholders such as learners, teachers and general public are made aware of the clarifications in testing and assessment practices. At the very same, test developers are promoted to get to grips with decision-makers from their institutions and ministries. Henceforth, decision-makers are made aware of the fact that there are both good and bad practices in testing and assessment, which leads the path to the improvement of assessment systems, and enhancement in the quality of the on-going assessment practices.

A review of recent studies on the utilization of the EALTA guidelines of good practice in language testing and assessment. As seen above, the EALTA guidelines are the arteries ending with a short-cut key to accomplish the goals set by the EALTA. In this vein, the use of the EALTA Guidelines has been consolidated by successive researches conducted in the field so far. However, there is a scarcity of empirical studies when it comes to practicality.

To probe into, Alderson and Banerjee (2008) have devised a questionnaire to the Aviation English test providers within the scope of considerations for test development in national or institutional testing units or centers. Alderson (2010) has made a report on Aviation English Testing regarding the guidelines set by the EALTA. Erickson and Figueras (2010) have noted a large-scale dissemination of the EALTA guidelines. To add more, De Jong and Zheng (2011) have conducted a case study applying the Guidelines on Pearson Test of English (PTE) Academic. As a result, the Guidelines together with codes of practice and ethical considerations are offered to be used to "frame a validity study" (Alderson, 2010, p. 63). Similarly, Kavaklı and Arslan (2017) have conducted a practical case study on the application of the EALTA Guidelines in the Foreign Language Proficiency Test administered in Turkey (YDS). As a result, they have reported that YDS could not correspond with the sub-criteria set by the EALTA Guidelines although the EALTA promotes value-added language testing and assessment implementations. Furthermore, the national school-leaving examination of Austria has been changed

from a teacher-designed form to a more standardized one for many of the foreign languages, such as English, French, Italian and Spanish in a project team's perspective (Spöttl, Kremmel, Holzknicht & Alderson, 2016). Therefore, the achievements and challenges have been evaluated in virtue of the EALTA Guidelines to raise awareness and adopt a new approach into language testing and assessment. Recently, Toncheva, Zlateva and John (2017) have conducted a study on developing a methodology in order to assess deck officers' language proficiency in Maritime English. Herein, they have applied the general principles of the EALTA Guidelines to create balance amidst test reliability, construct validity, authenticity and test usefulness.

Revising the literature, the author of this dissertation has shaped the study pursuant to eight major themes set by the EALTA Guidelines. Relatedly, it is reported that linking language tests and/or exams to the CEFR is a labyrinthical endeavor, taking years to be developed and resulted as a blending of a scientific approach, peer review and expertise. The EALTA, herein, makes clarification for the stakeholders who are involved in the linkage process such as teachers, policy makers, students and the general public addressing test quality and its impact, test developers' familiarization with the CEFR, the analysis of test content and test specifications, the procedures of standardization, the benchmarking of the performances to the CEFR, the publicly available reports on the linking process, the evidence of test reliability and validity, and the scheme of quality standards in language testing and assessment.

Consequently, a panorama to the review of literature in relation with the EALTA Guidelines, and language testing and assessment is drawn above. Below, the guidelines of the Association of Language Testers in Europe are presented together with its objectives, standards and resources. Last but not least, a review of recent studies conducted on the utilization of the ALTE Code of Practice in language testing and assessment is also done.

The ALTE: Objectives, Standards and Resources

The ALTE strives for setting common standards for language testing and assessment practices, whereby supporting multilingualism for the preservation of cultural and linguistic diversity in Europe. Taking this as the starting point, the author

of this dissertation discusses the objectives, standards and resources in detail to frame the functioning of the ALTE. First of all, the administrative body of the ALTE is enlightened together with its main objectives. Following that, two major scopes of the ALTE to accomplish its primary objectives are underlined: 'setting standards' and 'sustaining diversity', which is closely related with the concept of multilingualism, mushroomed as a must to maintain diversity across Europe. Within the scope of setting standards, the Code of Practice, Minimum Standards, and Portfolios are examined in detail with its exemplifications. On the other hand, the last section, is composed of recent studies conducted on the utilization of the ALTE Code of Practice in language testing and assessment.

The main body and objectives. With the abolishment of the international barriers amidst European nations and the increase in global migration, multilingualism becomes the reality throughout the world. Therefore, leaning towards fairness and accuracy in language teaching and assessment blossoms as a must in practice. This is due to the fact that multilingualism not only brings along benefits for many different societies, but it also threatens some societal and political systems as it may jeopardize the survival of languages from smaller communities - even in the hometown. Concerning all these together, the ALTE was founded in 1989 by Cambridge and Salamanca Universities to meet the demand for a lucid approach in language testing and assessment practices.

With 34 members, 40 institutional and several hundred individual affiliates, the ALTE works for promoting multilingualism by 'setting standards' and 'maintaining diversity' in Europe representing the testing of 26 different languages (ALTE, 2012). The ALTE aims to set common standards for language testing and assessment, and supports multilingualism for the preservation of the cultural and linguistic enrichment of Europe. In this respect, test takers can have the opportunity to be qualified by means of fair and accurate assessment criteria recognized around the world. Bolstering transnational recognition of certification in languages, the ALTE enables test takers to make comparisons with the qualifications they get in other languages. In addition to these, the ALTE makes use of joint projects, the works of special interest groups, bi-annual meetings and conferences in order to promote mobility and accessibility throughout Europe. To fulfil the above stated aims, the ALTE has put forward a strategic plan for the years 2013-2016, concentrating mainly on three

main themes. Firstly, the participation is to be widened by means of engaging stakeholders who are involved in language testing and assessment. Secondly, the examinations are to be improved concerning the significance of the 'ALTE Quality Management System'. Thirdly, the promotion of cooperation and partnership is a need to endorse multilingualism within and beyond Europe.

The ALTE canalizes into two major scopes: setting standards and sustaining diversity. To probe into, the increase in international mobility has mushroomed the demand for transferable language qualifications. To meet this demand, the ALTE has set a compile of common standards embracing the overall language testing process for its members. This process includes test development, item writing, test administration and analysis, marking and grading, together with the reporting process of the results. Therefore, the members of the ALTE benefit from professional specifications which are previously devised and delivered by the Association itself.

In doing this, the ALTE applies for its own newly-introduced quality indicator, the 'ALTE Q-mark', by which member organizations check for the accessibility of quality standards. Herein, the profile of an exam is audited whether to meet all 17 minimum standards set by the Association within the scope of test construction, administration and logistics, marking and grading, test analysis, and communication with stakeholders. Accordingly, the findings are reported after a rigorous audit in order to award an exam by Q-mark. An exam, which is awarded by Q-mark, enables test takers and/or users to feel assured as the aforementioned exam is proved to be appropriate by the Association. On the purpose of ensuring appropriateness in implementation, the ALTE makes use of guidelines for language testing, namely the 'Code of Practice', the 'Minimum Standards' embracing the criteria for effective language testing, and the 'Portfolios' for the promotion of independent learning environment and self-evaluation.

As mentioned previously, the ALTE canalizes into two major scopes. The ways of setting standards by the ALTE have been explained in detail above. Besides, in order to sustain diversity, the Association probes into the main theme of multilingualism. In 2013, a seminar on supporting multilingualism through language assessment was held by the European Parliament along with the ALTE and Cambridge English Assessment. The leitmotif of the seminar was the importance of

underpinning all languages in Europe, paving the way towards multilingualism to be able to withstand the current economic crisis. Bringing the term multilingualism to the fore, this seminar formed the grounds of assuring quality of language assessment practices, and putting a value on various languages by means of sustaining diversity throughout Europe.

In search of a common ground which binds all together, the ALTE respects diversity in all practices regarding language assessment. Therefore, its members represent different kind of languages spoken across Europe, even the less-widely spoken ones such as Welsh, Slovenian and Basque. To add more, with the help of regular seminars and conferences held, the Association works in cooperation with other European Institutions (e.g. CoE) to sustain diversity within Europe. These organizations are backed up through publications on multilingualism, as well. All of the ALTE's publications and other resources are easily accessible at its own website in 27 different languages.

The code of practice. In an attempt to define standards in examinations for current and future ALTE members, the Code of Practice was introduced in 1994. In this context, the Code of Practice states the liabilities of language examination providers, users and takers in which preliminary objectives are scrutinized under the heading of comparability of the quality so as to frame common levels of proficiency. These main users of language examinations are identified by the Code of Practice in detail. The developers are defined as those who construct and administer examinations. The users are labelled as those who select examinations and make decisions on the examination results. On the other hand, takers are listed as the candidates who take examinations.

Herein, it is to be noted that the roles of the developers and users do overlap one into another as they both set policies for the development process and makes decisions on the interpretation of the results. Therefore, the development and administration processes have a direct impact on takers, whose rights are also safeguarded by the Code of Practice. These rights include four core areas: developing examinations, interpreting the results, striving for fairness and informing examination takers. In doing these, the Code of Practice makes use of two types of responsibilities imputed on members at one side and examination users at the other side.

In order to develop examinations, its members are expected to define the characteristics of each exam such as the purpose, population(s), measurement concepts, examination development and administration process. In doing this, the members are to provide some representative samples regarding the examination intended to be used, and to clarify the concept of content and skill testing criteria. The procedure is to ensure the appropriateness of the examination to the target group within the scope of ethnic and linguistic backgrounds. On the other hand, as examination users are waiting for examination developers to provide information about the examinations, they are in a position to choose the one(s) which cater(s) their needs and seem(s) appropriate.

For the interpretation of the results, the members are to guarantee prompt and comprehensible reports, on which the examination takers' performances are stated clearly. Therefore, passing marks and/or grades are defined for each candidate. If there is not any certain marking and/or grading scheme, related information is to be provided to the takers in a reasonable way so as to prevent any misinterpretation and misuse of the results. On the other hand, examination users are expected merely to interpret the scores in a correct way.

Fairness is another topic to be considered by both members and examination users. Fairness is to be enabled within the scope of race, ethnicity, gender, and any other handicapping situation. In practice, the examination materials are to be reviewed and revised by the members in order to avoid potential misunderstandings. The language and content of the material are to jugulate insensitiveness. The differences which are intended to be assessed should be bounded solely to the performances of the takers, albeit not on race, gender or ethnic background. The test administration process for the candidates with a handicapped situation should be handled feasibly by means of available modifications, as well. On the other hand, fairness for users are framed within the appropriateness regarding the candidates' different backgrounds.

As a last step, members of the Association are to inform examination users and takers on selecting and implementing appropriate examinations. The candidates are to be informed equally about the coverage of the examinations such as task formats, strategies to be conducted, rubrics in use and other related instructions. The rights of the candidates should be framed within the concepts of

copyright issues and the release of the examination results. The same obligatory factors are valid for examination users if they are directly involved in the communication process with the candidates, as well.

The minimum standards. As mentioned previously, the Association has set standards to establish quality profiles in exams. Herein, five main points are considered: ‘test construction’, ‘administration and logistics’, ‘marking and grading’, ‘test analysis’, and ‘communication with stakeholders’. Accordingly, the entire examination process from A to Z is certified with the theme of fairness for all candidates.

To begin with, test construction as one of the minimum standards of the ALTE warrants the theoretical construct that the examination is grounded upon. The purpose, population(s), contextual use of the examination, required review and revision processes are described within. In order to create more consistent and stable peripheries, parallel examinations are probed and compared. If there is any linkage to a reference system such as the CEFR, grounds of alignment are claimed, as well.

For the administration and logistics part, the regulations are in the grip of examination centers, by whom transparency is maintained. Essential security systems are ensured for administering examinations and transporting the examination papers in tow. If there are any support systems like web and/or phone services for administration process, confidentiality of those systems is also guaranteed. All the candidates are informed about their rights, and data protection procedure legislated by law. The candidates with special needs are also considered if they are in need of support during the administration process.

Within the scope of marking and grading, accuracy and reliability are the ultimate aims to be accomplished throughout the process. Documentation can be a way of explanation about data collection process. Therefore, the analysis of estimated reliability and raters’ scores on writing and speaking performances can be highlighted. To add more, data should be collected on the candidates’ achievement scores in order to get rid of any influential factors underneath the success and/or failure such as country of origin, L1, age, gender and ethnic origin. Furthermore, item-level data, which can be listed as computing the reliability,

discrimination, difficulty and standard errors, are collected from the representative sample of the candidates and analyzed.

It is a crystal-clear fact that the examination administered also requires communication with stakeholders in order to announce the results. In that, examination centers, and the candidates themselves can learn the results more promptly and clearly. This communication process does not only embrace the announcement of the results. It also includes sharing information on the context, purpose and use of the examination. If enabled, this information sharing process helps the stakeholders to interpret the results more prominently to be used appropriately.

The portfolios. The ALTE has a set of guidelines for language testing, which are labeled as the Code of Practice, Minimum Standards and Portfolios. Herein, the portfolios are in use for the promotion of independence in learning and self-evaluation. In this context, the ALTE exploits two types of portfolios: the ELP and EAQUALS- ALTE Portfolio.

To enable integration and mobility across Europe, the ELP has popped up as a project, which provides learners with the recording of formal and informal learning experiences. Therefore, learners can keep track of both results of classroom practices such as examinations, and other learning experiences developed outside of the classroom. Accordingly, they are equipped with an overall picture of abilities, which is, at the very same, recognized in all parts of Europe. Supporting life-long learning, the ELP encourages individuals to update their recordings at regular intervals. In this sense, the ELP provides informal learning environment, as well. Motivating individuals to having a part in Europe, the ELP helps everyone to pursue his/her progress, even partially, in several languages. Thereby, it backs up pluri-lingualism and pluri-culturalism.

As elaborated previously, the ELP is composed of three subdivisions: 'language passport', 'language biography' and 'the dossier'. In sum, language passport is mostly used for external purposes with its reporting function up to six languages, including native language. Language biography records language learning experiences whereas the dossier comprises of sample materials which support other parts, namely language passport and biography. On the other hand,

in order to reflect diversity and comparability amidst examinations, ALTE members have developed an ALTE version of the ELP together with the EAQUALS. With the aim to act as a tool for enhancing language learning and having success in the end, the Portfolio gives responsibility to the individual to be aware of his/her own learning capacity. It also provides ALTE members to record their results and compare their ALTE examinations with others, creating a fair link to the CEFR. The EAQUALS-ALTE Portfolio is now available in seven languages, and preparation in other languages are also in progress. The accessibility to the Portfolio is gained by both paper-based and electronic versions with a guided pathway to its usage in tow.

A review of recent studies on the utilization of the ALTE code of practice in language testing and assessment. Above, the ALTE is probed in detail with its objectives, standards and resources. However, there is a scarcity of empirical studies merely focusing on the ALTE Code of Practice, instead there are studies conducted on the utilization of the Framework and the ALTE Minimum Standards for the alignment of the language tests.

Accordingly, Taylor (2006) has delved into the key elements to frame the varieties of English used within language tests, and their contributions to the community to make a better understanding of language variation. Herein, she has come up with some implications for language assessment, which are directly linked to the standards suggested by the ALTE. Additionally, Choi (2008) has provided an overview of the EFL context in Korea by framing the impact of standardized EFL tests. While exploring the nature of EFL tests in use, he has applied standards set by the ALTE, as well. In their study on using electronic portfolios for second language assessment, Cummins and Davesne (2009) have presented the American adaptations of the ELP, labelled as the 'Global Language Portfolio' and 'LinguaFolio', as the subsidiaries to be used with the Framework. Herein, they have applied the standards of the ALTE as a reference to reinforce the theoretical background. Correlatively, Xi (2010) has suggested the standards of the ALTE as one of the European criteria to enable test fairness, and to set priorities.

Consequently, a panorama to the review of literature in relation with the ALTE Code of Practice, minimum standards, and language testing and assessment is drawn above. Below, the guidelines of International Language Testing Association are presented together with its objectives, specifications and resources. Last but not

least, a review of recent studies conducted on the utilization of the ILTA Guidelines for Practice in language testing and assessment is also marked.

ILTA: The Objectives, Specifications and Resources

ILTA is a group of well-respected scholars and practitioners from the field of language testing and assessment, who are additionally internationally-recognized. Taking this as the starting point, the author of this dissertation discusses the objectives, specifications and resources in detail to frame the functioning of ILTA. First of all, the administrative body of ILTA is touched on together with its primary objectives. Following that, two major resources applied by ILTA to accomplish its primary objectives are underscored, named as the Code of Ethics and Guidelines for Practice. Within the scope of the Code of Ethics, nine fundamentals on ILTA members' ought-to-does and ought-to-not-does are identified. Correlatively, the test developers' and users' responsibilities at one side, and those of test takers at the other side are explained in two parts within the scope of the Guidelines for Practice. Following these, other resource types proposed by ILTA are briefly mentioned. As the last section, recent studies conducted on the utilization of the ILTA Guidelines for Practice in language testing and assessment are touched upon.

The body and primary objectives. As above mentioned, ILTA is a group of internationally-recognized and well-respected scholars and practitioners from the field of language testing and assessment. This group tries to define what it means to be a language tester with the purpose to promote the development of language testing practices in the world. Accordingly, ILTA aims to stimulate a notable achievement in the field of language testing through the dissemination of information amidst its members. In order to achieve these objectives, ILTA applies for two major resources: the 'Code of Ethics', and 'Guidelines for Practice', both of which are mentioned in detail below.

The code of ethics. ILTA bolsters ethical standards in language testing by means of the Code of Ethics, adopted at the annual ILTA meeting in Vancouver in 2000. The Code of Ethics is constituted by principles, benchmarking ethical behaviors of all language testers. These principles are framed within the scope of justice, respect for autonomy and civil society, beneficence and non-maleficence. In this sense, the Code of Ethics pinpoints 9 fundamentals. Accordingly, ILTA provides

its members with their 'ought-to-do'es and 'ought-to-not-do'es by identifying the complexities and exceptions in the implementation of these principles. Herein, the Code of Ethics relies on the morals and ideals of the profession as a response to the needs and changes of the profession. Therefore, failure to follow these principles by the members leads to the withdrawal of ILTA membership upon the advice of the ILTA Ethics Committee.

To elaborate, first principle probes into the concept of respect for test takers' dignity and privacy, which is a must for all language testers. The language testers should respect the needs and values of their test takers'. The language testers cannot influence test takers on the issues of ideology, politics and spiritual matters. Any act of discrimination is forbidden. The second principle deals with language testers' keeping information by their own professional capacity so as to share it with test takers confidentially. Language testers are required to respect the rights of their test takers. They are also obliged to safeguard the information gathered as a result of tester and test-taker relationship. These are documented as the professional duties of language testers to maintain confidentiality.

The third principle indicates that language testers are to abide by all ethical principles illustrated by both national and international standards if they are going to conduct any research activity, trial and/or experiment. As language testing involves the participation of human as the sample, research on the field of language testing is to follow the general principles of an academic inquiry. The research is to conform to the highest scientific and ethical standards. Consent of the all subjects should be free and flexible to withdraw from when not desired. The results of the research should be reported accurately and clearly. However, the identification of the participants who are enrolled in the study should not be announced when the research reports are published. The fourth principle guarantees the rights of the language testers so as not to misuse their professional knowledge against the interests of their test takers. Additionally, the fifth principle backs up language testers on the enhancement of their professional knowledge, and sharing it with other language professionals. Language testers are to keep themselves up-to-date with the latest developments and novelties in the field, and apply them for the goodness of their test takers. Henceforth, language testers are expected to have a seat at professional conferences, regular workshops, annual meetings and/or

seminars, and to follow the publications in recognized journals related to their profession.

The sixth principle requires language testers to share the responsibility of endorsing integrity among colleagues in the language testing profession. Therefore, a sense of trust blossoms mutually through exchanging opinions and viewpoints in order to develop and exercise norms for the sake of society. In the event of unprofessionalism conducted by any of a colleague, language testers are expected to report the situation to the authorities with utmost seriousness. The seventh principle encumbers language testers with the societal role of quality improvement in language testing and assessment practices. It is, herein, to be noted that they should remember their role as educators at one side, and their role as citizens at the other side. Accordingly, language testers should share knowledge and expertise, and advise language testing services for the enhancement of quality. In doing these, language testers are to refrain from self-promotion and derogation of their colleagues.

The eighth principle entails that language testers are to be aware of their responsibilities to the test takers, stakeholders and overall society. Language testers are to accurately report the results for the sake of universities, schools, related departments, professional bodies, and the like. Language testers should also comply with the testing requirements of the society in which they work, even when they are not pleased with. Lastly, the ninth principle requires language testers to consider both short- and long-term potential effects of their practices on stakeholders. Therefore, language testers are assumed to contemplate ethical considerations of the projects, which are rendered to them. Following a deep evaluation, language testers should report possible consequences of these projects, as well. If there is any professionally unacceptable situation, then they are to negotiate the situation with fellow language testers to find a fair ground.

The guidelines for practice. Besides the Code of Ethics, ILTA also proposes the Guidelines for Practice, whose draft version was firstly introduced at the ILTA meeting held in Ottawa in 2005. Following this, the circulation among ILTA members yielded the development and adoption of it at another ILTA meeting in Barcelona in 2007. The final revised version was found fully appropriate in 2010. Composed of two main parts, the ILTA Guidelines for Practice offer basic

considerations for good testing practice in all situations such as “responsibilities of test designers and test writers, obligations of institutions preparing or administering high stakes examinations, obligations of those preparing and administering publicly available tests, responsibilities of users of test results, special considerations, and rights and responsibilities of test takers” (ILTA, 2007, p. 1-8). In epitome, Part A is concerned with the test developers’ and users’ liabilities whereas Part B deals with the test takers’ rights and liabilities.

To broach Part A, basic considerations deals with the key assumptions to be covered for good testing in all situations. In this vein, test developers should develop an understanding for each construct of the test. For instance, the purpose of the test should be clearly stated. In order to make solid inferences, the test constructs are to measure what they are supposed to measure so that validity is supplied. The test results should be consistent and comparable so that reliability is provided. In doing these, test designers and test writers have some responsibilities, as well. In a proper test design, the specifications and statements created by the test developers and/or designers should refer to the intended purpose of the test explicitly. Before the pre-testing stage, each test task should be edited to report if there is any malfunctioning within. Before administering the test, marking schemes should be prepared. The scoring stage should involve inter- and intra-rater reliability calculations, which are also expected to be published. Besides, the test results should be interpreted accurately by all test takers. As indicated, test designers should point out the test tasks in detail, and safely keep the test materials with special care. In all practices, all test takers should be treated in the same way to ensure equality.

Moreover, there are some obligations for institutions while preparing and/or administering high stakes examinations. These institutions can be exemplified as schools, certification bodies, colleges etc. In order to enroll in such kind of institutions, test takers should apply for high stakes examinations. For these kind of examinations, the test preparation stage should be firstly depended upon the language testing theory which is currently in use. For those who are non-native speakers of the language being tested, someone with a high level of proficiency in the aforementioned language is to be employed to check the items written. All test takers should be provided with satisfactory information about the procedure. The results gained at the end should be announced correctly and put in the data-base

after a continuous quality control analysis. It is also to be noted that if there is more than one form, inter-form reliability is to be calculated and published, as well.

For those who prepare and administer tests which are publicly available, it is compulsory to signal the sample group which is targeted to be tested. In order to prevent any misleading claim, a handbook, in which reliability and validity scores, test purpose, measurement concepts, scoring criteria, marking scheme and relevant information are penned, should be published and publicly announced for test takers. Besides, for those who are the users of test results, the responsibilities are defined as making fair decisions on the results, clearly interpret the results for the goodness of test takers, highlighting the limitations of the test results before decisions are based on, being well-prepared to put evidence on the accuracy of the decisions made, and bearing the standard error of measurement (SEM) in mind before making a decision on the results gained.

Furthermore, there are some special considerations involving three types of testing: norm-referenced, criterion-referenced and computer adaptive testing. In norm-referenced testing, the features of the sample group should be reported as appropriate or not, in order to set standards for comparability, before the test is normed on. In criterion-referenced testing, the appropriateness of the criterion is at the helm of the experts in the field. For the calculation of reliability and validity scores, basic correlation analysis is not found suitable. Therefore, proper methods should be conducted accordingly. In computer adaptive testing, the sample group is expected to be rather larger to assure the cohesion of the Item Response Theory (IRT) calculations. Additionally, test takers and other stakeholders are to be mindful of the distinction between computer adaptive tests, and traditional paper and pencil tests.

On the other hand, Part B deals with the rights and responsibilities of the test takers. They have the right to be respected without any discrimination on the issues of gender, ethnicity, religion, age or any other personal characteristics. Professional standards are to be used in testing process, and all test takers are to be informed about the process beforehand. Test takers should also be informed about the characteristics of the test: whether taking the test is optional or not. The results of the tests should be kept in good care to ensure confidentiality under law, and by the Code of Ethics for the sake of test takers' privacy. Besides, test takers have the

responsibility to be aware of their own rights and liabilities as the test takers. Therefore, they are also expected to treat others with respect. They are to know about the place where the test will be conducted, and the duration when the test is required to start and finish. During the testing process, the test takers are expected to follow the instructions given. In case of a comprehension difficulty experienced during the testing process, the test takers can consult to an examiner and inform him/her about the situation. If the test taker is in need of a special care due to his/her physical condition and illness, s/he is again expected to inform an examiner in advance of testing so as not to be influenced by external factors on his/her performance. In a respectful manner, all test takers are welcomed to present any concern about the testing process by and large.

Other resources. As above mentioned in detail, ILTA has some resources which define the key principles underneath. In addition to these, ILTA makes use of some other resources available for those who are involved in language testing process. In this context, ILTA provides a bibliography which is composed of dissertations written about language testing and assessment within years. This bibliography acts as a service to students of language testing in order to ensure that their studies can be as comprehensive as possible, which is updated regularly and presented online. On the other hand, ILTA announces a compile of research activities such as conferences, seminars and webinars under the heading of upcoming events. All language testing conferences are given in detail by means of a calendar of events. To add more, for those who are interested in research studies conducted in the field of language testing and assessment, there is a list of academic journals, as well. By those means, ILTA tries to blossom a common understanding among those who are studying in the field of language testing and assessment worldwide.

A review of recent studies on the utilization of the ILTA guidelines for practice in language testing and assessment. Above, ILTA is probed in detail with its objectives, specifications and resources. However, there is a scarcity of empirical studies separately and merely focusing on the ILTA Guidelines for Practice, instead there are studies conducted on the utilization of the Framework and the ILTA Code of Ethics in designing language tests.

Accordingly, Shohamy (2001) asserts that language testers should be accountable to give information on the test results. Therefore, ethical principles of utmost importance. Correlatively, Davies (2008) has examined the changes in language testing textbooks in English. As, more recently, the concepts of validity and fairness get on the stage, the ILTA Code of Ethics is applied as an explicit declaration of ethical principles. Besides, Brown and Bailey (2008) have investigated the features of language testing courses within the scope of course characteristics, instructors and learners. In doing this, they have marked the ILTA Guidelines for Practice while reporting the differences and similarities between the 1996 and 2007 results both qualitatively and quantitatively. In relation with classroom assessment, Mendoza and Arandia (2009) have probed into the perceptions of teachers about language assessment in Colombia. The results of their study have indicated that more importance is to be given to the training of teachers about language testing and assessment. Besides, they have reported that the ethical principles of ILTA should be taken into account for the training of teachers. Last but not least, Xi (2010) has proposed an approach to guide practitioners who are interested in fairness on language testing and assessment practices in order to provide a way of integrating fairness and setting priorities for it. In doing this, she has probed into TOEFL® iBT™ to demonstrate how fairness in testing could be established and backed up in a validity argument by applying the ILTA Code of Ethics as one of the fundamentals of sustaining ethics in testing and assessment practices.

Consequently, a panorama to the review of literature in relation with the ILTA Guidelines for Practice, and language testing and assessment is drawn above. Below, the guidelines of the Association for Educational Assessment- Europe are presented together with its purposes, core elements and guiding principles. Last but not least, a review of recent studies conducted on the utilization of the AEA-Europe's Framework in language testing and assessment is also scrutinized.

The AEA- Europe: Purpose, Core Elements and Guiding Principles

The AEA- Europe is a platform where developments in educational assessment in Europe are discussed, leading cooperation between individuals and organizations. Taking this as the starting point, the author of this dissertation

discusses the purpose, core elements, guiding principles and instrument within the context of the European Framework of Standards for Educational Assessment. Firstly, a general description of the AEA- Europe is touched upon together its main purposes targeted at. Following that, the core elements in liaison with this Framework are explained one by one, namely (a) 'goal construct'; (b) 'nature of evidence of tasks'; (c) 'gathering evidence'; (d) 'capturing outcomes'; (e) 'decision-making'; (f) 'interpreting and reporting results'; (g) 'evaluation and next iteration'. Concomitantly, the guiding principles of this Framework are listed and elaborated one by one, which are basically to (a) 'focus on educational assessment'; (b) 'fit for a European environment'; (c) 'emphasize ethics, fairness and the rights of the individual'; (d) 'address essential quality concerns of validity, reliability and impact on stakeholders'; (e) 'support learning, test development and review'. Following these, the instrument is briefly mentioned to show the implementation of the Framework. As the last section, a review of recent studies regarding the utilization of the AEA- Europe's Framework is touched upon.

The definition and purposes. The AEA- Europe serves as a platform where developments within the scope of educational assessment within Europe are discussed to cherish collaboration between individuals and related organizations. Therefore, it promotes educational assessment practices together with academic, professional and vocational contexts. In doing this, the AEA- Europe organizes conferences to bring the ones who are interested in educational and occupational assessment together. Besides, it provides the bare bones of a research to foster joint projects across Europe. Engaging individuals, agencies and organizations in a myriad of activities to improve assessment practices and products in Europe, the AEA- Europe develops an understanding for the impact of those practices in any educational environment.

The AEA- Europe functions as a Council whose body is occupied with different types of committees. These can be labeled as scientific program committee, conference organizing committee, publications committee, professional development committee and other ad hoc committees. To support the running, the AEA- Europe benefits from two major committees: professional development and publications committees. Professional development committee supports continuous professional development at the core in order to create opportunities for its members

to keep themselves up-to-date with the latest developments in assessment. On the other hand, the latter, publications committee, strives for creating an environment in which individuals maintain a professional medium of communication through internet by means of e-newsletters, LinkedIn groups and social media channels. Among the LinkedIn groups, there is a doctoral network group of the AEA- Europe, in which doctoral students as future leaders of the field are able to communicate with each other, and share career matters together.

To accomplish above mentioned purposes, the AEA- Europe has developed the 'European Framework of Standards for Educational Assessment' (AEA- Europe, 2012). In a word, this framework offers standards to foster transparency for both users and educational authorities by benchmarking the on-going system of standards for the enhancement of further assessment processes. Among these, the core elements, guiding principles and instrument are highlighted in detail. Each sub-component is explained below with their constituents in tow.

The core elements. With the intention of providing an instrument for educational authorities, test providers and score users to compare their assessment practices, the European Framework of Standards for Educational Assessment has flourished as an evidence for above-mentioned types of audience. As a subcomponent of the Framework, the core elements spring from assessment development cycle, constituted by seven standard requirements following one another: "(1) defining the goal; (2) identifying the nature of evidence and of tasks; (3) gathering evidence; (4) capturing outcomes; (5) decision-making; (6) interpreting and reporting results; (7) evaluation and next iteration" (AEA-Europe, 2012, p. 9).

To clarify, while defining the goal of assessment, the construct, group and function should enter into the process. As a first step, what the test is going to measure such as knowledge, aptitudes, skills and the like should be clearly stated. The test takers as a group should be settled within the aspects of age, occupation, educational level etc. Additionally, the inferences which are expected to be drawn from the results, and the intended users of those results are to be defined. The strengths and weaknesses undergone throughout the assessment process should be illuminated for further amendments. As a second step, the nature of evidence and of tasks should be clearly identified. In this context, the intended behaviors of the test takers are to be elicited by tasks. The design of the assessment procedure

should sufficiently represent the content which is covered by knowledge, skills and other attributes, and the setting in which the assessment is going to take place. Task types should all be up to the mark, and prevent from discrimination amidst test takers in order to maximize construct-relevant variance. However, the comparability of the test results to sort out more competent ones than the others is fundamental to make a review on the tasks and related materials better. Herein, it should be noted that sufficient amount of evidence is needed in order to make valid inferences.

While gathering evidence as a third step, the main concepts pop up as administration and logistics. In the collecting of evidence, no security leaks are allowed. If the test is administered online or through internet, the assessment procedures should be identified properly for the sake of test taker's identity. One more to add, the background variables which have an effect on the outcomes should be used as a reference for the evaluation of the outcomes. The purpose of the fourth step, namely capturing outcomes, assures the validity of the gathered outcomes. Herein, scoring and/or marking should be separated from grading. The instruments which are used in the process should be evaluated in quality before the measurement. If the scorer and/or marker is human, intra- and inter-rater reliability scores are to be considered to set rater-agreement on common standards. Outcomes of any test should be collected feasibly to facilitate the process.

Decision-making as the fifth step requires a combination of the test outcomes. At this point, the type of combination and the rationale behind it are to be clearly specified. The performance standards or norms in use such as group-referenced, domain-referenced and/or criterion-referenced, should be selected in harmony with the testing criteria and test characteristics. It should be noted that when there is a norm, there are to be cut-off scores. Therefore, the decision behind that certain cut-off scores should be brightly declared. For standardized assessments at high-stakes, the test difficulty and score comparability is to be investigated in order to ensure fairness among test takers, as well. Before evaluation, the results are to be interpreted and reported accurately. The proper interpretation of the results should be handled with written policies or guidelines to assure data protection. The reporting process should be framed within the variables of format, content and timing. Moreover, confidentiality is to be enabled by related documented policies while informing the candidates about the scores.

The last step, evaluation and next iteration, embraces the use of the results for further cases. Herein, the concept of next iteration might be comprised of either the developing a new form of assessment, or improving the already existing one. The adaptation of an assessment is also included within the context of next iteration. In this sense, three basic elements should be focused upon: 'evaluation of technical aspects', 'evaluation of the usefulness' and 'evaluation of the impact of the assessment'. The very first one, evaluation of the technical aspects of the assessment nestles psychometric analyses. On the other hand, the evaluation study of the usefulness flags the efficiency and practicality of the administration process. Besides, the impact of an assessment is to be concerned while shaping learner behaviors and addressing teaching professionals together with other community members. In that, all assessment practices have a strike on both learning and educational outcomes even including families of the learners. The degree of impact sharpens and widens if the stakes get much higher, though.

The guiding principles. The Framework is grounded upon five major guiding principles. As the Framework goes at educational assessment, it, therefore, goes hand in hand with the European standards. It also highlights ethics in order to ensure individual's rights through fairness. It focuses on practicality, validity and impact on stakeholders as the essential quality concerns. Yet, it supports not only learning, but also decision-making and test development processes.

To elaborate, the very first major guiding principle is that the Framework focuses on educational assessment. Herein, the assessment types which support learning are addressed. The testing situations are generally composed of the assessment of formal learning such as summative school-based assessment, vocational assessment, performance assessment etc. However, innovative types are also considered as the new forms of assessment. As the second guiding principle, the Framework is to fit for a European environment. In that, the AEA-Europe contributes to the development of quality in educational assessment with its European perspective in a world-wide interest. With its integrative function, the Framework gathers the on-going traditions in assessment and new forms of approaches together. It also backs up variety in cultural and educational contexts in order to enlighten what is underneath the concept of fitness-for-purpose. Therefore,

local definitions are also brought to the agenda while disseminating quality in educational assessment.

The third principle suggests that the Framework underscores ethics, fairness and the rights of the individual. As the main beneficiaries of the Framework, the individuals are the prominent elements of the assessment process. At that point, ethical considerations are also given due weight in order to guarantee test takers' rights. The assessment process is not inscribed merely to the test administrators and developers, albeit to test takers. As the fourth principle, the Framework addresses essential quality aspects such as validity, practicality and impact on stakeholders. As already known, the aforementioned quality aspects are the cornerstones of a professional assessment. Therefore, the results of a test should be meaningful and useful for every test taker. At the very same, the results should reflect certain degree of credibility relying on fundamental assessment principles.

The fifth principle ascertains that the Framework supports learning, decision-making, test development and program review processes. If well-devised, any assessment procedure will surely enhance learning. However, learning is effected negatively if this procedure is designed haphazardly and/or poorly. Providing feedback is essential for both decision-makers and program reviewers in order to enhance the quality of educational assessment, and to evaluate programs. In doing so, the Framework follows the assessment development cycle, which is basically composed of standard requirements clarified within aforementioned seven core elements, methods of implementation and possible evidences.

The instrument. The instrument enables all three levels of the Framework, which are labeled as 'standard requirements', 'methods' and 'samples of evidence', to work together. Herein, the standard requirements are defined previously within the scope of core elements, which are rather directive. An assessment procedure should meet all those standard requirements, and address them. Additionally, the elements are generic and illustrative, albeit not specific and prescriptive. However, the elements are composed of the methods and examples of evidence for each standard requirement; therefore, they are to be described by means of observations and verifications, though.

A review of recent studies on the utilization of the AEA- Europe's framework in language testing and assessment. Above, the AEA- Europe is probed in detail with its purposes, core elements and guiding principles. However, there is a scarcity of empirical studies merely focusing on the AEA- Europe's Framework, instead there are studies conducted on the utilization of the CEFR, which are laced with the principles of the AEA- Europe in enhancing the viewpoints towards educational assessment.

Accordingly, Jones and Saville (2014) have highlighted the importance of Learning Oriented Assessment (LOA) with a systemic view. LOA is actually grounded upon the socio-cognitive model of language learning propounded by the Framework. It is noted that such an approach has either been “explicitly or implicitly defined in opposition to traditional externally set and assessed large scale formal examinations” (Davison & Leung, 2009, p. 395). They have embarked on the language assessment resulting from classroom interactions, and suggested the adoption of this approach abiding by the guiding principles of the AEA- Europe in order to develop the current educational assessment practices. In a similar vein, it is suggested by Halbherr, Schlienger and Piendl (2014) that assessment practices should be molded in reply to globalization around the world; therefore, assessment for a digital world is to be revised and re-arranged in accordance with the Framework.

Similarly, as educational assessment has some essential quality concerns not only for learning but also for decision-making and test development processes, teacher assessment literacy is supposed to be enhanced consequently. Herein, DeLuca, LaPointe-McEwan and Luhanga (2015) have made a review of international standards and measures, in which they have touched upon the guiding principles of the AEA- Europe, as well. As one of the core professional requirements across all educational systems, the standards for assessment literacy adopted in five countries, namely Australia, Canada, New Zealand, the UK and USA have been probed with special interest on the measures developed after 1990. Henceforth, they have drawn a general frame of changes in the assessment practices over time and across different countries, which are all English-speaking ones. Correlatively, Wools (2015) has developed an evaluation system of validity in order to enhance the quality of educational assessment by means of the results of a design-based

project. Within, the theoretical principles and designing tenets are correlated with the guiding principles of the AEA- Europe in order to develop a prototype for validity.

To note more, the Annual Conferences of the AEA- Europe are embellished with various studies on the enhancement of educational assessment practices. Amidst the recent ones, Van Nijlen and Janssen (2014) have touched upon national assessments to measure the 21st century skills, with special reference to that of information processing. Besides, Zumbo (2015) has explored the consequences and side effects of an ecological model of testing (Hubley & Zumbo, 2011), in which the assessment is considered something *in vivo* rather than *in vitro*. Herein, Jones and Saville (2009) has suggested the Framework as a model for learning, and as an instrument of harmonization. One more to note, Jones (2007) has contributed to the relationship between assessment and National Languages Strategy. A framework to accredit language proficiency, labelled as 'The Languages Ladder', has been investigated by means of a system developed by 'Cambridge Assessment: Asset Languages'. In applying these, he has addressed the Framework in order to create opportunities for language assessment, and, herewith, to improve the quality of language assessment.

Chapter Summary

Above, the European standards in language testing and assessment are probed in detail. In doing this, each of the standards are elaborated separately with their sub-components. Accordingly, the Framework is explained through its origins and content together with the development of the RLDs, certification systems, ELP and Manuals in order to frame the language testing and assessment standards of the CEFR. Furthermore, the EALTA is highlighted with its mission, principles and main considerations. Herewith, the ALTE is explored within the concepts of its objectives, standards and resources by means of the 'Code of Practice', 'Minimum Standards' and 'Portfolios'. To note more, ILTA is scrutinized by its objectives, specifications and resources by means of the 'Code of Ethics', and the 'Guidelines for Practice' whereas the AEA- Europe is clarified through its core elements and guiding principles. Consequently, a panorama to the review of literature in relation with the utilization of each European standard above mentioned, and language testing and assessment is drawn. The following chapter is about the national

standards of language testing and assessment practices together with those of NFE in Turkey.



Chapter 3

The National Standards of Language Testing and Assessment in Turkey

Introduction

In this chapter, the national standards of language testing and assessment in Turkey are discussed. Accordingly, the very first part starts with a brief outline of the Turkish National Education System together with the changes undergone. Besides, the language testing and assessment practices of formal educational settings are mentioned highlighting the Measuring, Selection and Placement Center (ÖSYM) in Turkey. Additionally, those of non-formal educational settings are scrutinized together with the implementations of the Ministry of National Education (MoNE), and the Association of Private Educational Institutions and Study Centers in Turkey (ÖZ-KUR-DER). Other relevant units responsible for language testing and assessment across the country are also touched upon. As the last section, a brief chapter summary is given a place to draw a general picture of the language testing and assessment practices in Turkey.

The Turkish National Education System

Adopted as the basic premise with the foundation of the Turkish Republic, education has been considered as one of the most influential factors in the nation-building process across the country. Concomitantly, the principles of 'universality and equality' are inscribed within the Basic Principles of National Education by the Basic Law of National Education, which is also claiming that national education is organized around the demands of the society, and in parallel with the society's abilities, capabilities, interests and skills by the Articles 5 and 6 of the Basic Law No. 1739 legislated by the MoNE in June, 1973.

In other respects, the United Nations Educational, Scientific and Cultural Organization (UNESCO) has announced a report on a move towards life-long education (UNESCO, 1972). This has led to a tripartite categorization of the education systems taking life-long learning as the core element (Colardyn, 2002; La Belle, 1982). Just because formal education systems are more conservative to adapt the socio-economic changes around them swiftly, there occurs a point of departure which highlights the distinctions among formal, in-formal and non-formal

education around the world (Fordham, 1993). As a result, NFE has mushroomed as an educational force of the postmodern world, which develops into the worldwide educational industry (Romi & Schmida, 2009).

For the Turkish context, it is due to the radical changes in the general political environment of the late 1980s when the politics of education have undergone a gradual withdrawal of the state from education by creating new opportunities for the private sector (Demirer, 2015). Accordingly, the basic premises of national education in Turkey have also become inappropriate and inadequate to meet the demands of the current society, as many other countries finding it difficult to pay for the expansion of formal education. As the education has become more individualized, out-of-school education system has been enlightened more than before. The results of a study conducted by the CoHE on higher education pupils have demonstrated that the proportion of the learners who are enrolled in a private institution in order to meet their further learning needs is 71.8% (CoHE, 2007). In this context, it is reported by the Association of Private Educational Institutions and Study Centers in Turkey that the number of private institutions has reached up to 1.500 in 2011, which is approximately 600-750 million Turkish lira revenues. That is why the sudden change in the Turkish education system with the total closure of some of these private institutions, namely *dershanes* in Turkish context, or returning them to the Basic High Schools, has caused some problems (Dolgunsöz, 2016). However, the Association also emphasizes that the opening of new private courses by the municipalities and other non-governmental organizations has resulted in an unfair competition by operating in a wrongful way (ÖZ-KUR-DER, 2011). Supporting inequalities in education by such wrongful implementations, these private institutions have become more prevalent for those rushing in a competitive environment where success becomes hard to be accomplished (Silova, Budiene & Bray, 2006; Southgate, 2009).

In the light of these, the national education system of Turkey is constituted by two main sections: (1) formal education (hereafter FE); and (2) non-formal education (NFE) in tune with the Basic Law No. 1739 for National Education. FE refers to the schooling system composed of pre-school education, primary education, secondary education and higher education respectively. On the other hand, NFE is fed by all other educational activities organized outside and/or

alongside the FE. Therefore, NFE encompasses any organized educational activity based on an out-of-school system, operating separately, or as a part of a much broader activity in order to serve for recognizable learning objectives (Coombs, Prosser & Ahmed, 1973). To set a comparison between FE and NFE, attendance with pre-defined age limits is compulsory in FE; however, any citizen can benefit from NFE with no age limit. For FE, one needs to complete primary education in order to continue with secondary education. Contrary to FE, NFE does not require previous schooling to step in further education, as each phase is running independently from each other.

To note more, NFE is embellished with general and vocational technical programs. Herein, the institutes providing NFE could be listed as the practical arts schools, advanced technical schools, industrial practical arts schools, technical education centers, public education centers offering craftwork, literacy courses, tech-related courses and language-related courses, and apprenticeship training centers. In all cases, the Turkish MoNE holds the responsibility to run FE and NFE systems. However, the ratio of auditing is rather low for non-formal educational settings than formal educational settings. Not to mention, the testing and assessment practices of formal educational settings are conducted by the Measuring, Selection and Placement Center in Turkey whereas those of non-formal educational settings are held by the MoNE.

Language Testing and Assessment in Formal Educational Settings

The testing and assessment practices of formal educational settings are conducted by the Measuring, Selection and Placement Center in Turkey. The Measuring, Selection and Placement Center, known as ÖSYM, is the body which is responsible for orchestrating large-scale examinations on a national-level basis in Turkey (ÖSYM, 2013a). Amidst these examinations, the most commonly known is the Higher Educational Examination named as the Student Selection and Placement System, which is a standardized test used for the admission of high school students to the universities in Turkey. Besides, the Academic Personnel and Graduate Education Examination (ALES), Vertical Transfer Examination (DGS), Public Personnel Selection Examination (KPSS), The Turkish Medical Specialty Examination (TUS), the Examination for Foreign Students (YÖS) could be listed as

some of those large-scale examinations in Turkey, which are all administered by ÖSYM.

In addition to these, the language proficiency tests are also administered by ÖSYM in Turkey. Two different types of language proficiency tests were administered by ÖSYM until 2013, which are named as the Foreign Language Examination for Civil Servants (KPDS), and the Inter-University Foreign Language Examination (ÜDS) (Külekçi, 2016). With the arrival of the new language proficiency test, namely the Foreign Language Examination (YDS), ÖSYM becomes responsible for administering it to evaluate the foreign language skills of the test takers, who are generally civil servants, military personnel academics and graduate students. With its high-stake nature, the Foreign Language Examination administered in Turkey is expected to pin down some European standards. However, this examination is only accredited in Turkey, and is approved to be used merely within the country for further purposes.

To briefly mention, the examination, which has also been conducted electronically (i.e. e-YDS) since January, 2017 is administered every six months. The examination administered in spring term is conducted in more than twenty languages such as Chinese, German, French, Japanese, Spanish and the like. On the other hand, that of fall term is conducted merely in Arabic, German, English, French and Russian. Composed of 80 multiple-choice question items, the foreign language proficiency examination in Turkey is primarily engaged with grammar, translation, odd-one-out, reading comprehension, sentence completion and vocabulary parts. Herein, the test takers are given 180 minutes to complete the examination, each item of which is 1.25 points in scoring. The false answers are not separately eliminated from the correct answers, albeit just noted as 'false' out of the overall scores. It is, herein, to be noted that for the examinations conducted in Armenian, Chinese, Danish and Greek, solely translation questions, which are evaluated by an academic jury of ÖSYM, are asked, though. One more to note, ÖSYM adopts British English in language proficiency tests prepared (ÖSYM, 2016a), asserting that the resources applied for test development and design procedures are those of 'inner-circle countries' (Kachru, 1992). However, the test is solely composed of items measuring receptive language skills although the test items are considered as authentic and original (Akın, 2016).

For the evaluation of the results, ÖSYM exploits its own alignment system, in which A accounts for 90- 100 points, B accounts for 80- 89, C accounts for 70- 79, D accounts for 60- 69, and E accounts for 50- 59 (ÖSYM, 2016b). Besides, the scores above are also aligned with the proficiency levels defined by the Framework, albeit on a one-way recognized interpretation as the results could only be used for local purposes. Accordingly, A1 refers to 30- 44 points, A2 refers to 45- 59 points, B1 refers to 60- 74 points, B2 refers to 75- 94 points, C1 refers to 95- 99 points, and C2 refers to 100 points (ÖSYM, 2016c). Similarly, ÖSYM has its own equivalence tables for other language proficiency examinations (e.g. TOEFL® iBT™), which are declared by the CoHE, and are molded with the proficiency levels of the CEFR, as well (ÖSYM, 2016c).

Language Testing and Assessment in Non-Formal Educational Settings

MoNE is indirectly involved in the process of testing and assessment practices of the institutions serving for NFE in Turkey. In other words, the language certificate examination of the non-formal educational institutions is administered by MoNE in Turkey. Herein, the aforementioned institutions are expected to submit a petition to the Ministry after arranging the lists of all test takers. Following this, the foreign language certificate examinations are prepared by the English language teachers bounded to the District Directorate of National Education. There is no standard applied in test development process; however, the test items are prepared according to the curriculum followed. The language certificate examinations are held every two months on weekends. On Saturdays, the multiple-choice tests are conducted whereas oral proficiency examinations are sit on Sundays. There are totally 47 exam centers throughout the country, in which the test takers sit for the language certification exam. The successful ones are given a General English Certificate by the institution they are enrolled in.

The association of private educational institutions and study centers in Turkey: ÖZ-KUR-DER. ÖZ-KUR-DER is the Association of Private Educational Institutions and Study Centers in Turkey with a head office in Ankara, and 304 members in number around the country. The Association initially aims to meet the common demands of its members in social, economic, educational, cultural and professional arenas by fostering cooperation among its members. As an Association, ÖZ-KUR-DER endeavors for constituting a harmony for its members

abiding by the standards code related to private educational institutions. Besides, it struggles for building a systematic view for on-going educational practices, and for developing skills and knowledge. In doing this, ÖZ-KUR-DER organizes seminars, workshops or conferences in and outside of the country.

In order to enhance the quality of Turkish national education standards, the Association has interiorized some general principles such as honesty, respect to the law, professional competence, prestige, accountability, fairness and confidentiality. Herein, the Association adopts a fair approach abiding by the law in conducting the above-mentioned services. Besides, the Association organizes summer workshops to develop their members' professional competence and prestige. The Association is also principled in a non-conciliatory manner to pursue non-formal educational activities across the country. In doing these, the Association sticks to the confidentiality of the information gathered, and fairness in managing a relationship with its members.

Moreover, ÖZ-KUR-DER, together with its General Assembly, Board of Directors and Board of Auditors, contributes to the objectives of Turkish MoNE in many aspects. To exemplify, it promotes free courses for the children harbored by the charities and/or non-profit organizations. By establishing commissions, it supports designing course materials such as textbooks, and conducting analyses in the presence of the Board of Education and Discipline. It also tries to eliminate the so-called private institutions which are actually working 'under the counter' by auditing them, and informing the authorities concerned. In doing this, the Association forms a platform by receiving support from other associations, foundations and non-governmental organizations in order to accomplish a common purpose.

Apart from these, ÖZ-KUR-DER holds pilot tests partake of the central large-scale examinations in order to enhance the quality in testing and assessment practices. In a similar vein, the Association holds examinations together with the Vocational Qualifications Institution in line with the demands of the members and/or on behalf of the member institutions, and keeps an account of the results for further use. To note more, the Association has made a request to MoNE for providing all of

their learners with the ELP at the end of the courses in the private institutions rendering foreign language education.

However, there are some difficulties encountered by the Association in fulfilling the responsibilities aforementioned (ÖZ-KUR-DER, 2017). To exemplify, the education levels of the course takers are not defined by the modules of the Directorate General for Life-long Learning. Thus, the requirements of the curriculum aimed to be covered are not set on a standard basis, as an institution sets literacy as the main requirement to be enrolled in whereas the other sets primary school graduation to start. Besides, the duration of the courses is much longer than desired. Therefore, even the course takers who tend to keep up with their lacks in learning are to follow the curriculum set for the fresh starters. Indeed, the former group of learners are to be taken to placement tests in all kinds of institutions as soon as they are enrolled in.

With special reference to the private institutions rendering English language education, the teachers are selected from MoNE via the District Directorate of National Education. They are given permission by the Ministry to render 10 course hours of lecture in these private institutions. Additionally, ÖZ-KUR-DER also promotes the underachieving learners to take extra tuitions within the body of the Association until the implementation of the language certificate examination. Therefore, the English language teachers are to receive bachelor's degree at the very least to work in these private institutions. Besides, the teachers are also provided with in-service training programs held biyearly by the Association on a volunteer basis. However, the language certificate examinations prepared by the English language teachers of those private institutions are not penned in accordance with the European standards; henceforth, not applicable for further use. Therefore, the ratio of test takers is rather low compared to those taking the accredited examinations, which are internationally recognized in all educational arenas.

Other Relevant Units of Language Testing and Assessment

In addition to the Foreign Language Proficiency Examination administered by ÖSYM, a similar examination was also introduced in February, 2017 by the CoHE. Accordingly, Higher Education Institutions Foreign Language Examination

(YÖKDİL) is put into use by the protocol signed among Anadolu University, Ankara University and the CoHE (CoHE, 2017). The main difference between YDS and YÖKDİL is that YÖKDİL is the foreign language proficiency examination that addresses three types of audiences separately. The trilateral segmentation of the audiences is those of Social Sciences, Physical Sciences and Medical Sciences. Since then, it has been accepted as valid, and the results gained from it have been used as equivalent to those of the Foreign Language Examination. Hence, Anadolu University and Ankara University together with the CoHE serve as the relevant units of large-scale foreign language examinations in Turkey.

Moreover, there are also some other exam centers rendering services for those who are eager to take internationally accredited foreign language proficiency examinations. For the ones who are going to take TOEFL® and the International English Language Testing System (IELTS™), there are exam centers in Ankara, Istanbul and Izmir. On the other hand, TOEIC® is solely conducted by the local network office in Istanbul. Besides, the Pearson Test of English Academic (PTE Academic™) is applied in Adana, Ankara, Denizli, Diyarbakır, Erzurum, Istanbul and Kayseri in Turkey. In addition to the study abroad programs which require PTE Academic™ as the requirement for enrollment, there are also 8 universities in Turkey, which accept PTE Academic™, as well. Similarly, Cambridge Assessment English is another certificate examination applied to study abroad. There are 15 exam centers in total, which are located in Ankara, Antalya, Eskişehir, Fethiye, Istanbul and Izmir.

Chapter Summary

In this chapter, the national standards of language testing assessment in Turkey are discussed. In doing this, ÖSYM, MoNE, ÖZ-KUR-DER and other relevant units responsible for language testing and assessment practices in Turkey are scrutinized in order to draw a general picture of language testing and assessment practices in Turkey. Within other relevant units, the CoHE, MoNE and other exam centers are also mentioned together with the foreign language proficiency examinations administered across the country. A distinction is made between formal and non-formal educational settings together with the implementations of language testing and assessment.

Chapter 4

Methodology

Introduction

In this part, the methodology of this dissertation is presented in detail. Information on the research design in general, the population as the sample size and setting, and materials used as data collection instruments are touched upon. Moreover, the research design exploited within the procedure is explained step by step. Data gathered through several components during the data collection process are uncovered in detail. Therefore, data analysis procedure including statistical techniques laced with demographic information, descriptives and related testing of assumptions are clearly shown. And, finally, ethical considerations are explained for the confidentiality of the data collection process.

Research Design

In this study, which is based on the mixed methods research design, which “as a method, focuses on collecting, analyzing, and mixing both quantitative and qualitative data in a single study or series of studies” with “its central premise to use of quantitative and qualitative approaches in combination” in order to “provide a better understanding of research problems than either approach alone” (Creswell & Plano Clark, 2007, p. 5). Therefore, both qualitative and quantitative data were collected in order to arrive at an understanding of the on-going testing and assessment practices of three institutionalized private English language schools offering education in their branches in all of the major cities in Turkey. In the study, especially European standards for testing and assessment were taken into consideration. Accordingly, a questionnaire composed of ‘5-point-Likert-type’ response items (1, strongly disagree; 2, disagree; 3, not sure; 4, agree; 5, strongly agree) has been conducted with the aim of gathering relevant quantitative data. Besides, semi-structured interview sessions, as a part of qualitative data, were led by the researcher within the scope of general information about the institution; the running of the on-going testing and assessment practices laced with numeric data on the number of teachers, test (-item) developers, students and the like; and the difficulties and problems encountered in the implementation of the testing and assessment practices together with the recommendations for further improvement,

which paves the way for mixed methods research design. Accordingly, the quantitative data facilitated the testing of the hypotheses on a positivist paradigm by means of measuring variables and empirical data (Marczyk, DeMatteo & Festinger, 2005; Sarantakos, 2005). On the other hand, with the help of qualitative data, the researcher could get multifaceted realities of the participants within a natural context on a constructivist paradigm (Candy, 1991). Therefore, human behaviors, which were “fluid, dynamic and changing over time and place” (Johnson & Christensen, 2012, p. 35) were taken into consideration through reciprocal constructions of perceptions and thinking.

In this context, the qualitative data were gathered from the directors of these private institutions and the director of ÖZ-KUR-DER whereas the quantitative data were gathered from teachers who were also working as test (-item) developers at the same private institutions. The quantitative data were analyzed by Statistical Package for Social Sciences (SPSS) Version 23.0. On the other hand, the qualitative data were analyzed by the constant-comparison analysis method (CCA) mentioned in detail below. The findings were given numerically within tables, which were followed by the results discussed in tow.

Participants and Setting

Before delving into the details concerning the participants, it is to be enlightened that three major non-formal private institutions serving as English language schools in Turkey as the source of subjects were selected for this study. For the selection process, the primary concern was to cooperate with the most prominent courses which were renowned for quality in learning English in Turkey with the highest course attendee capacity and with the highest number of branches in Turkey in order to enable the generalizability of the results. Besides, those selected 3 private institutions were the members of ÖZ-KUR-DER, as well. Taking the recommendations of the director of ÖZ-KUR-DER on the selection process, the researcher also adopted ‘convenience sampling’ (Dörnyei, 2007; Nunan, 1992) as a technique concerning the fact that participants could be more convenient for accessibility by the researcher.

In the light of these, the data were collected in the fall term of the academic year 2016-2017 with the participation of 40 English language teachers (12 male and

28 female participants) recruited from aforementioned 3 English language schools, whose name were kept anonymous for the confidentiality of the results; therefore, labelled as A, B and C. The English teachers participated in the study were each counted as 11, 19 and 10 from the above labelled English language schools respectively. The participants' age range ranked from 18-25 (N= 27) and 26-35 (N= 12) to 36-45 (N= 1). When their years of experience were considered, teachers mostly had the experience of less than five years (N= 32) which was followed by 5 to 9 years (N= 6) and more than 14 years (N= 2) respectively. One more to note, all of the participants were both English language teachers and test (-item) developers at the private institutions they were working. The table given below summarizes the demographic information about the participants:

Table 1

Overall Demographic Information of the Participants

		N	Percentage %
Institution	A	11	27.5%
	B	19	47.5%
	C	10	25.0%
Gender	Male	12	30.0%
	Female	28	70.0%
Age	18-25	27	67.5%
	26-35	12	30.0%
	36-45	1	2.5%
Years of Experience	less than 5	32	80.0%
	5-9	6	15.0%
	more than 14	2	5.0%
Occupational Field	teacher	40	100.0%
	test (-item) developer	40	100.0%
	Total N	40	100.0%

General information on the institution A. In order to get healthy information on the institutions, the director of each was asked about the total number of branches in Turkey, the number of English language teachers working within, approximate number of students enrolled, and other explanatory information. Accordingly, the institution A was reported by its director to have a sum of 64 branches in Turkey. It had an amount of approximately 650 English language

teachers in total working either part- or full time. The number of students was reported to be nearly 300 for each branch.

Besides, the institution A was noted by its director to offer English language courses for the levels from A1 to C2, as depicted by the CEFR. A2 level was noted as 'elementary'. There were 6 classes in total; each of which was composed of 80 hours of lecture in sum. Each course was equal to a level. The courses were given in two alternates: for 3 days in a week throughout 9 months; or 2 days in a week throughout 15 months. Based on the teaching of fundamental language skills, the program was constituted by the courses rendered through skills-based language teaching.

To elaborate, the institution A selected as the sample for this study was composed of 11 English language teachers who were also working as test (-item) developers at the selected private institution. Of those, 7 were female (63.6%), and 4 were male (36.4%) with the age range of 18-25 (N= 7; P= 63.6%) and 26-35 (N= 4; P= 36.4%). Additionally, they had the years of teaching experience ranging from less than five years (N= 8; P= 72.7%) and from five to nine years (N= 2; P= 18.2%) to fourteen years and above (N=1; P= 9.1%). respectively. The table given below summarizes the demographic information about the participants from the private institution A:

Table 2

Demographic Information on the Institution A

		N	Percentage %
Institution	A	11	27.5%(of total)
Gender	Male	4	36.4%
	Female	7	63.6%
Age	18-25	7	63.6%
	26-35	4	36.4%
Years of Experience	less than 5	8	72.7%
	5-9	2	18.2%
	more than 14	1	9.1%
Occupational Field	teacher	11	100.0%
	test (-item) developer	11	100.0%
Total N		11	100.0%

General information on the institution B. In order to get healthy information on the institutions, the director of each was asked about the total number of branches in Turkey, the number of English language teachers working within, approximate number of students enrolled, and other explanatory information. Accordingly, the institution B was reported by its director to have a sum of 153 branches in Turkey. It had an amount of approximately 1.500 English language teachers in total working either part- or full time. The number of students was reported to be between 1.700 to 2.000 per year solely for the branch that took part in this study. Herein, it was also noted by the director that the smaller the branches in Turkey became, the lesser the number of students enrolled in.

Besides, the institution B was noted by its director to offer English language courses for the levels from A1 to C2, as depicted by the CEFR. The courses were rendered through skills-based language teaching. Each course was composed of a standard of 80 hours of lecture. The courses were given in two alternates: for 3 days in a week throughout 9 months; or 2 days in a week throughout 15 months. As mentioned, the program was based on the teaching of fundamental language skills. Therefore, if a student quit the course when s/he was at B1 level, that student was expected to start the course from the very beginning again, which was actually A1 level.

The institution B was also noted to provide social activities for students by the clubs opened. There were either unlimited clubs such as vocabulary clubs, grammar clubs, and speaking clubs, or carrier clubs such as how to write a CV, prepare yourself for interviews in English, etc. The private institution B did not concentrate solely on students. The English language teachers working within the private institution B were also provided with in-service training facilities such as how to teach English to the speakers of other languages (TESOL), how to use body language effectively, the art of rhetoric, and the like.

Furthermore, the institution B selected as the sample for this study was composed of 19 English language teachers who were also working as test (-item) developers at the selected private institution. Of those, 16 were female (84.2%), and 3 were male (15.8%) with the age range of 18-25 (N= 17; P= 89.5%) and 26-35 (N= 2; P= 10.5%). Besides, they all had less than five years of teaching experience (N=

19; P= 100%). The table given below summarizes the demographic information about the participants from the private institution B:

Table 3

Demographic Information on the Institution B

		N	Percentage %
Institution	B	19	47.5%(of total)
Gender	Male	3	15.8%
	Female	16	84.2%
Age	18-25	17	89.5%
	26-35	2	10.5%
Years of Experience	less than 5	19	100.0%
Occupational Field	teacher	19	100.0%
	test (-item)	19	100.0%
	developer		
Total N		19	100.0%

General information on the institution C. In order to get healthy information on the institutions, the director of each was asked about the total number of branches in Turkey, the number of English language teachers working within, approximate number of students enrolled, and other explanatory information. Accordingly, the institution C had a sum of 12 branches in Turkey, 2 of which were reported to be under construction by its director himself. It had an amount of 250 English language teachers in total working either part- or full time. It was also reported to have 10.000 students per year all around Turkey.

Besides, the institution C was noted by its director to offer English language courses for the levels from A1 to C1, as depicted by the CEFR. Each course took a time of 1 year for all levels. Hence, a student enrolled for A1 took a one-year course, which was actually the same with another student enrolled for C1. A one-year program was composed of 104 hours of lecture on main courses, which were also laced with additional courses noted as 48 hours of lecture per month. A sum of 150 hours of lecture constituted the essential program of one class. Moreover, weekly course hours were reported to be shifting between 8 and 16.

Moreover, the institution C selected as the sample for this study was composed of 10 English language teachers who were also working as test (-item) developers at the selected private institution. Of those, 5 were female (50%), and 5 were male (50%) with the age range of 18-25 (N= 3; P= 30%), 26-35 (N= 6; P= 60%) and 36-45 (N=1; P= 10%). Additionally, they had the years of teaching experience ranging from less than five years (N= 5; P= 50%) and from five to nine years (N= 4; P= 40%) to fourteen years and above (N=1; P= 10%). The table given below summarizes the demographic information about the participants from the private institution C:

Table 4

Demographic Information on the Institution C

		N	Percentage %
Institution	C	10	25.0%(of total)
Gender	Male	5	50.0%
	Female	5	50.0%
Age	18-25	3	30.0%
	26-35	6	60.0%
	36-45	1	10.0%
Years of Experience	less than 5	5	50.0%
	5-9	4	40.0%
	more than 14	1	10.0%
Occupational Field	teacher	10	100.0%
	test (-item) developer	10	1000%
	Total N	10	100.0%

Instruments

In this section, the instruments used to collect data for this study are presented.

Instrument 1: A questionnaire on the European standards for establishing quality profiles in exams. In order to uncover the testing and assessment practices of aforementioned private institutions rendering English language education in Turkey, some European standards for establishing quality profiles in exams were listed considering the guidelines proposed by the EALTA; the Manual recommended by the CEFR; Guidelines for Practice introduced by the

ILTA; the Code of Practice ascertained by the ALTE; and general educational assessment guidelines set by the AEA-EUROPE.

Accordingly, a questionnaire composed of 87 items on a 5-point Likert-type response basis was administered for this study. The first section of the questionnaire aimed to collect demographic information about the sample group such as gender, age, years of teaching experience and occupational field. The second section of the questionnaire was composed of 87 minimum standards for establishing quality profiles in exams. These standards were aligned with the criteria set by above-mentioned European guidelines, and were arranged in the format of a '5-point Likert-type scale', in which 'Strongly Disagree' was the lowest possible rating and 'Strongly Agree' was that of highest. The test items were all molded into a table adjacent to the cells next to each test item. During the arrangement process, the wording of the questionnaire was slightly modified as the aforementioned European guidelines put forward the requirements to be followed in related testing and assessment practices. More precisely, instead of 'The tests should require ...' pattern, 'The tests in use require ...' pattern was employed in the wording of each test item. Herewith, the participants were asked to read each statement carefully and circle the number in the cells (from 1 to 5) which was the best descriptor of their own opinions, ensuring that there was not any correct or false answer, and all of the information that could identify them would remain confidential.

The minimum standards were set in liaison with the aforementioned European guidelines. However, they were not gathered together, evaluated and exploited by researchers all at once. Therefore, in order to check the internal consistency of the scale used, a reliability analysis was conducted. As a prior step, negatively worded items were estimated as three, and were noted as item no. 24, 25 and 87 respectively, which were all coded reversely. Then, overall Cronbach's Alpha level for the instrument was evaluated for the context in which the present study was conducted:

Table 5

Reliability Co-Efficiency of the Data Collection Instrument

Reliability Statistics	
Cronbach's Alpha	Number of Items
.952	87

As seen in the table above, Alpha reliability co-efficient of the data collection instrument was .952, indicating that the reliability of the data collection instrument was considered to be strong, as a scale in Social Sciences was expected to have the reliability score of .70 Cronbach's Alpha at least to be fair. Additionally, the internal consistency was also checked by split-half reliability. It was yielded by the split-half reliability analysis that Cronbach's Alpha for the first part was .933 (r_1 for 44 items) whereas that of the second part was calculated as .909 (r_2 for 43 items). As noted, there was a high internal consistency within items and no problematic data entry was identified. Correlatively, the Alpha value was also checked if there was any increase thanks to any item deletion. However, as there was no substantial increase in Alpha value through reliability analysis, none of the items was eliminated.

As noted previously, the minimum standards set within the questionnaire were categorized in terms of pre-determined European guidelines purported by the CEFR, ALTE, ILTA, EALTA and AEA-EUROPE. As the standards identified by the ALTE, ILTA, EALTA and AEA-EUROPE were also certified and confirmed by the CEFR, Manual was not separately included within the questionnaire not to recap the same items again and again, albeit intertwined together. The outline of these standards could be seen in table given below:

Table 6

An Outline of the Minimum Standards in the Questionnaire

Section(s)	Sub-Section(s)	Number of Items
1. The ALTE Code of Practice and Minimum Standards	1.a. Test Construction	10 (Item No. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
	1.b. Administration & Logistics	6 (Item No. 24, 29, 30, 31, 32, 33)
	1.c. Marking & Grading	7 (Item No. 22, 25, 34, 35, 36, 37, 43)
	1.d. Test Analysis	10 (Item No. 11, 12, 13, 14, 15, 16, 17, 18, 19, 20)
	1.e. Communication with Stakeholders	3 (Item No. 38, 39, 40)
	1.f. Test Production	5 (Item No. 21, 23, 26, 27, 28)
	1.g. Item Writing	2 (Item No. 41, 42)
2. The EALTA Guidelines for Good Practice in	2.a. Quality Control and Test Analyses	5 (Item No. 44, 45, 83, 84, 85)
	2.b. Review and Washback	2 (Item No. 46, 87)

Language Testing and Assessment	2.c. Linkage to the CEFR	3 (Item No. 47, 48, 86)
	2.d. Test Design and Item Writing	1 (Item No. 58)
3. The ILTA Guidelines for Practice	3.a. Responsibilities of the Test Designers and Test Writers	6 (Item No. 49, 50, 51, 52, 53, 54)
	3.b. Responsibilities of the Test Takers	3 (Item No. 55, 56, 57)
4. The AEA-EUROPE Standards for Educational Assessment	4.a. Guiding Principles	19 (Item No. 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77)
	4.b. Instrument/ Identifying the Nature of Evidence, Tasks and Test Types	5 (Item No. 78, 79, 80, 81, 82)
TOTAL	15 Sub-Sections	87 Items

The outline given above was constituted concerning the order of the standards within the questionnaire. Besides, the data gathered by the questionnaire from the teachers and test (-item) developers were laced with semi-structured interviews with the directors of the institutions assigned, and with that of ÖZ-KUR-DER.

Instrument 2: A form of semi-structured interview sessions with the directors. In addition to quantitative data gathered by the questionnaire which was mentioned in detail above, the directors of the institutions were met and invited to the semi-structured interview sessions led by the researcher to get qualitative data. Every single session lasted for 15-20 minutes with each director. The sessions were conducted face-to-face on a volunteer basis, pursuant to the appointments arranged beforehand.

In essence, the director of ÖZ-KUR-DER was primarily visited for an interview on request. This meeting with the director of ÖZ-KUR-DER was organized in order to get general information about the private institutions and study centers serving for NFE in Turkey. The semi-structured interview session with him wheeled around the frame of the lack of the European standards in testing and assessment practices, the qualifications of the teachers working at private institutions, an approximate number of English language courses across the country, the deficiencies of the on-going system in testing and assessment practices, and some practical solutions and recommendations offered in the sequel. Upon the interview with the director of ÖZ-KUR-DER, the most prominent and widely-known three English language schools were selected, also bearing the capacity of course

attendee enrolled in mind. In the selection process, it was also considered that the selected private institutions were all the members of ÖZ-KUR-DER, as well.

Besides, the directors were initially asked about some general information about the running of the institution: the number of branches, teachers, students, course hours, types of testing and assessment practices, the European standards adopted, course duration and proficiency levels, and some other numeric data laced with recommendations. Moreover, 6 open-ended questions were asked to get more detailed information on testing and assessment practices conducted within the institution. These questions could be listed as given below:

1. Please provide some information on the testing and assessment practices conducted within your institution.
2. Are these practices aligned with any European standards? If yes, please provide some information about those standards.
3. Please provide some information about the instruments and the criteria set for testing and assessment practices.
4. Please provide some information on the difficulties and problems mostly encountered in testing and assessment practices.
5. Please provide some recommendations in order to enhance the on-going testing and assessment practices within your institution.
6. Please provide some recommendations in order to enhance the on-going testing and assessment practices across the country.

Procedure

During the study, the gap in the literature was found out, and the topic was determined before the information providers for the aforementioned topic were probed into. After the overall study plan was drawn, and the preparations were done, the data gathered for this study was collected in the fall term of the academic year 2016-2017.

First of all, the director of ÖZ-KUR-DER was primarily visited for an interview on request. This meeting was held with the director of ÖZ-KUR-DER, and was organized in order to get some general information about the private institutions and study centers of English language teaching and/or learning serving for NFE in

Turkey. The semi-structured interview session with him wheeled around the frame of the lack of the European standards in testing and assessment practices, the qualifications of the teachers working at private institutions, an approximate number of English language courses across the country, the deficiencies of the on-going system in testing and assessment practices, and some practical solutions and recommendations were offered in the sequel.

Upon the interview with the director of ÖZ-KUR-DER, the most prominent and widely-known three English language schools were selected, bearing the capacity of course attendee enrolled in mind. These selected three English language courses were visited and their teachers who were also test (-item) developers were appointed as the participants for this study. Then, they sat for filling in the questionnaire under the supervision of the researcher and the director of the institution. The sample group who took the questionnaire was composed of 40 English language teachers who were also working for test office as test (-item) developers. Following that, the directors of the aforementioned English language schools sat for the semi-structured interview sessions in person, which were all led by the researcher upon request. Considering the date of the appointments with each director, the directors at the private institutions were visited one by one. Each semi-structured interview sessions lasted for 15-20 minutes with each of the directors.

Data Analysis

Following data collection process by convenience sampling, the raw data were taken to analysis by aparring quantitative data at one side and that of qualitative one at the other side. For quantitative data, statistical procedures were employed via SPSS Version 23.0 after entering all the valid data in. On the other hand, the data gathered qualitatively, which were noted as the directors' reports both from ÖZ-KUR-DER and the institutions, were analyzed through constant-comparison analysis method. The selection of each statistical technique primarily depended upon accuracy and precision in essence.

The data gathered was taken to analysis with the identification of the demographic information first. Means, standard deviations and frequencies were noted for each test item as a part of descriptive statistics in order to spot "...general tendencies in the data and the overall spread of the scores" (Dörnyei, 2007, p. 213).

The mean scores gathered were ranked from the highest to the lowest in order to distinguish the most positive and more negative items assessed. Through descriptive statistics, each of the items were summarized enabling comparisons across the institutions selected, enabling researcher to compare the relative weightings of the exploitation of the European standards previously defined amidst the institutions.

When the normality is enabled, the Pearson Product-Moment Correlation Coefficient is over there to measure of the strength of a linear association between variables. The Pearson Correlation Coefficient is categorized as 1, perfect; .70- .90, strong; .40- .60, moderate; .10- .30, weak; and 0, zero (Dancey & Reidy, 2004). For this study, the Alpha reliability co-efficient of the data collection instrument was .952, indicating that the reliability of the data collection instrument was considered to be strong, as a scale in Social Sciences was expected to have the reliability score of .70 Cronbach's Alpha at least to be fair. Additionally, the internal consistency was also checked by split-half reliability. It was yielded by the split-half reliability analysis that Cronbach's Alpha for the first part was .933 (r_1 for 44 items) whereas that of the second part was calculated as .909 (r_2 for 43 items). As noted, there was a high internal consistency within items and no problematic data entry was identified.

Regarding qualitative data, the key features of a quantitative data analysis are put aside as it contradicts with the nature of the data gathered, and the general paradigm of qualitative data analysis (Mills, 2003). Therefore, internal and external validity together with the internal and external reliability are taken into consideration. Ensuring the internal validity, the literature review process was carefully followed in order to draw a conceptual frame beneath the semi-structured interview forms. Transforming the directors' interview reports into written forms, the researcher also asked each of the directors to control these written forms of texts, and to confirm that the forms were accurately typed down. Supporting transparency, the researcher had the emerging data codes checked by two other independent raters in order to build consensus. Additionally, the data analysis procedure was penned down by the researcher comprehensively to ensure external validity through comparability and transferability. Bounded to the internal reliability, the researcher reported all the findings without adding any other personal interpretation and generalization. With the help of two independent raters, the codes were also cross-checked to reach an

agreement. As a last step to enable external validity, the raw data gathered and codes emerged were kept by the researcher in order to provide opportunity for any further inquiry.

Besides, the semi-structured interview sessions with the directors of selected private institutions and ÖZ-KUR-DER was conducted in the first language of the director, which was Turkish. Therefore, after the sessions, the researcher translated the original version into the target language, which was English. With the help of back-translation method, two independent raters translated this version into the original language with no prior knowledge of the original content, enabling the researcher to consult with the translators to detect any discrepancies (Marin & Marin, 1991). In order to prevent the translated instrument skewed one-way and to reduce “human factor as each inquirer had his/her own unique final destination just like a scientific two-edged sword” (Platton, 2015, p. 433), those independent raters were selected concerning the fact that they had different background of knowledge, expertise and world view, albeit proficient in the target language. Herein, one-way translation method, which was deemed as the most unreliable method of all translation methods, was not applied since it was solely depended upon the knowledge of the individual translator (Erkut et al., 1999).

As a procedure, the analysis of the semi-structured interview reports of the directors followed a constant-comparison analysis method (Bogdan & Biklen, 2003). The constant comparison analysis method pursues a very similar way to the grounded theory approach, in which researchers come up with an emergent fit; therefore, they adjust the category to fit the data, albeit do not go for data to link with a pre-determined category (Taber, 2000). Herein, the constant-comparison analysis method encompasses a process of reducing the data gathered by means of constant recoding (Glaser & Strauss, 1967). Thus, the procedure is broken down into steps starting with the comparison between the already existing incidents, which is further pursued by the comparisons between concepts and incidents. Elliott and Jordan (2010) states that “... it is through the process of comparing concept to incident that the researcher can check to see if further incidents fit with the newly developed concepts and, in so doing, ensure that the concepts are capable of accounting for all related incidents in the data” (p. 34-35).

Based on this, the researcher designated codes to each line directly in the margins of the interview reports, associating entries with codes with similar meanings into a new category. This process continued for each of the remaining reports of the directors. Following a reiterative angle, codes from the first report were transferred to the second one, and those of the second report were carried over to the third one. This procedure made it possible to create thematic trends across the institutions, and the self-reports of their directors through reunification.

Ethical Considerations

The researcher contacted each of the directors and all of the English language teachers who were working at these private institutions in person, and gave them required information about the research and the ethical issues both orally and in written way. Besides, it was assured by the researcher that any information that could identify participants' names would remain confidential. It was also warranted that the information in this study would solely be used for research purposes and in ways that would not reveal the identity information of the participants. Finally, it was ensured by the researcher that the contribution of the participants might not render any personal benefit but might help to improve CEFR oriented testing and assessment practices of English language schools as non-formal educational settings in Turkey; thus, all of the teachers who were also working as test (-item) developers at their institutions accepted to be a part of this study as a participant on a volunteer basis.

Chapter 5

Findings and Results

Introduction

In this part, answers to the research questions and emerging sub-research questions are discussed and elaborated in detail. The findings of the data analysis are presented in tow. Basically, research questions which are aimed to be answered are singly presented.

Results of the Data Analysis

In this section, the results gathered are discussed separately for each private institution in relation to the research questions.

Do the testing and assessment practices of non-formal English language schools in Turkey comply with the criteria designated by the EALTA? The EALTA seeks for answers to those who are mingling with three types of considerations in testing and assessment practices: (a) considerations for teacher pre-service and in-service training in testing and assessment; (b) considerations for classroom testing and assessment; (c) considerations for test development in national and institutional testing units or centers. Taking these into consideration, those who are involved in the test development process on a national and an institutional basis are suggested with considerations probed into below.

To note beforehand, the items (N= 11) in the questionnaire regarding the 'EALTA Guidelines for Good Practice in Language Testing and Assessment' (EALTA, 2006) were numbered as 44, 45, 46, 47, 48, 58, 83, 84, 85, 86 and 87. The items were categorized into 4 groups. These groups were named by the EALTA itself as quality control and test analyses, review and washback, linkage to the CEFR, and test design and item writing. Considerations for test development in national or institutional testing units or centers involve test purpose and specification together with test design and item writing. However, as the ALTE Code of Practice also covered the same process under the heading of test construction and production in the questionnaire, the considerations set by the EALTA was not covered again. The items in the questionnaire could be listed as below:

Table 7

Questionnaire Items by the Guidelines of the EALTA

Section(s)	Sub-Section(s)	Item(s)
The EALTA Guidelines for Good Practice in Language Testing and Assessment	1. Quality Control and Test Analyses (5 items)	Item No. 44: The equivalence between different versions of the tests (e.g. year by year) are verified. Item No. 45: The actions to improve the quality of teaching and learning are taken after the implementation of each test. Item No. 83: The tests are piloted before they are administered to the target population. Item No. 84: The results are scored via automated scoring machines. Item No. 85: The results are scored via human scoring.
	2. Review & Washback (2 items)	Item No. 46: The test items keep pace with changes in the current ELT curriculum. Item No. 87: Traditional assessment practices are in use for test takers.
	3. Linkage to the CEFR (3 items)	Item No. 47: There is a publicly available report on the linking process between tests in use and the Reference Supplement, such as the CEFR. Item No. 48: As a part of the linkage to the CEFR, the tests correspond to the procedures recommended in the Manual and Reference Supplement. Item No. 86: Test takers are provided with contemporary self-assessment tools such as the European Language Portfolio (ELP).
	4. Test Design and Item Writing (1 item)	Item No. 58: Test item writers are trained before test administration.
TOTAL	4 Sub-sections	11 Items

A sum of 11 items, which were above listed in detail, was taken to frequency analysis through descriptive statistics one by one. To add more, for each item, the participants' answers from 3 institutions were estimated and reported singly.

Accordingly, the first main consideration of the EALTA, namely quality control and test analyses, was composed of 5 core items (item no. 44, 45, 83, 84 and 85). Quality control and test analyses involved equivalence between different versions of the test, the actions to improve the quality of on-going practices, piloting process and scoring procedure through previously set analyses. Each item was probed and described one by one to give detailed information on the estimations gathered.

Secondarily, the practices nestling review and washback were checked with the help of 2 items (item no. 46 and 87). It was asked whether the test items kept changes in the current ELT curriculum, which was also controlled by the test item questioning the types of techniques used in on-going testing and assessment practices: traditional vs. contemporary. Additionally, linkage to an external reference system; herein, the CEFR, was probed through 3 items (item no. 47, 48 and 86). Within the compass of these items, it was asked whether there was any publicly available report on the linking process between the tests in use and Reference Supplement, such as the CEFR. It was also asked whether the tests in use corresponded to the procedures recommended in the Manual and Reference Supplement ascertained by the Framework itself as a part of the linkage to the CEFR. Besides, the participants were asked whether the test takers were provided with contemporary self-assessment tools such as the ELP.

Last but not least, test design and item writing was composed of 1 item (item no. 58). Within, it was asked whether test item writers were trained before test administration. To note, as the other European guidelines for good practice in testing and assessment practices were comprised of many other items on test design and item writing, the same items were not taken into statistical analysis repeatedly for the reliability of the results. Before delving into details, table below given embodied the overall estimations regarding the exploitation of the EALTA Guidelines by selected private institutions. Means, standard deviations and standard errors of mean were given for each item elaborately.

Table 8

The Exploitation of the EALTA Guidelines by Selected Private Institutions

Section(s)	Item(s)	N	Mean	Std. Error of Mean	Std. Deviation
1. Quality Control and Test Analyses	Item No. 44	40	3.80	.119	.757
	Item No. 45	40	3.95	.123	.782
	Item No. 83	40	3.80	.165	1.04
	Item No. 84	40	3.40	.182	1.15
	Item No. 85	40	3.93	.126	.797
2. Review and Washback	Item No. 46	40	3.83	.107	.675
	Item No. 87	40	2.25	.099	.630
3. Linkage to the CEFR	Item No. 47	40	3.75	.117	.742
	Item No. 48	40	3.90	.106	.672

3. Linkage to the CEFR	Item No. 86	40	3.98	.116	.733
4. Test Design and Item Writing	Item No. 58	40	3.75	.159	1.01
TOTAL	4 Sub-sections/ 11 Items	40			

In the light of these, it could be stipulated that some actions were taken in order to improve the quality of teaching and learning after the implementation of each test with the highest mean score of all ($M= 3.95$; $SD= .78$) within the scope of quality control and analyses. It was followed by human-scoring, indicating that test results were, at the same time, scored by the participants themselves in the selected private institutions ($M= 3.93$; $SD= .79$). To some extent, the equivalence between different versions of the tests was pursued and verified on a pre-defined timely basis ($M= 3.80$; $SD= .75$). Alike, before the tests were administered to the target population, they were noted to be piloted with the mean score of 3.80/ 5.00 ($SD= 1.04$). Validating the result that the test results were still analyzed by human-scoring, the utilization of automated scoring machines in marking and grading was spotted to have the lowest mean score ($M= 3.40$; $SD= 1.15$), meaning that automated scoring machines were not marked as the only instrument of marking and grading; therefore, human scoring was still in use in some cases.

With a view to review and washback, the test items in use were stipulated to keep pace with changes in the current ELT curriculum with the estimated mean score of 3.83/ 5.00 ($SD= .67$). It was also noted as traditional assessment practices were still in use for test takers, even if just a smidgen ($M= 2.25$; $SD= .63$). Linkage to the CEFR was checked by three items which yielded the results that the test takers were provided with contemporary self-assessment tools such as the ELP with the highest mean score ($M= 3.98$; $SD= .73$) out of three. It was followed by the item purporting that the tests in use corresponded to the procedures as asserted in the Manual and Reference Supplement as a part of the linkage to the CEFR ($M= 3.90$; $SD= .67$). Furthermore, it was stipulated by the participants of this study that there was a publicly available report on the linking process between the tests in use and Reference Supplement, such as the CEFR with the mean score of 3.75/ 5.00 ($SD= .74$). Last but not least, for the sub-section of test design and item writing, the mean

score was estimated as 3.75/ 5.00 (SD= 1.01), indicating that test item writers were somehow trained before test administration.

Keeping these in mind, each sub-section was analyzed separately for each of the selected private institutions. The results were elaborated in detail, and the tables for each were given one by one. At first, an overall estimation regarding the results gained from all of the private institutions were checked and reported together. Following that, the results of each private institution were checked and reported separately by means of frequencies and percentages given within tables. With these in mind, the table below showed the overall results in a statistical order before delving into the results of each private institution in detail. Each item was reported underneath singly, and the overall estimations were supported by their implications.

Table 9

The Implementation of the EALTA Guidelines by Selected Private Institutions

The Implementation of the EALTA Guidelines of Good Practice in Language Testing and Assessment		Strongly Disagree	Disagree	Not Sure	Agree	Strongly Agree	TOTAL
1. Item No. 44: The equivalence between different versions of the tests (e.g. year by year) are verified.	f	0	1	13	19	7	40
	%	0.0	2.5	32.5	47.5	17.5	100
2. Item No. 45: The actions to improve the quality of teaching and learning are taken after the implementation of each test.	f	0	0	13	16	11	40
	%	0.0	0.0	32.5	40.0	27.5	100
3. Item No. 83: The tests are piloted before they are administered to the target population.	f	2	1	11	15	11	40
	%	5.0	2.5	27.5	37.5	27.5	100
4. Item No. 84: The results are scored via automated scoring machines.	f	4	4	9	18	5	40
	%	10.0	10.0	22.5	45.0	12.5	100
5. Item No. 85: The results are scored via human scoring.	f	0	1	11	18	10	40
	%	0.0	2.5	27.5	45.0	25.0	100
6. Item No. 46: The test items keep pace with changes in the current ELT curriculum.	f	0	2	7	27	4	40
	%	0.0	5.0	17.5	67.5	10.0	100
7. Item No. 87: Traditional assessment practices are in use for test takers.	f	0	2	8	28	2	40
	%	0.0	5.0	20.0	70.0	5.0	100
8. Item No. 47: There is a publicly available report on the linking process between tests in use and Reference Supplement, such as the CEFR.	f	0	2	3	32	3	40
	%	0.0	5.0	7.5	80.0	7.5	100

9. Item No. 48: As a part of the linkage to the CEFR, the tests correspond to the procedures recommended in the Manual and Reference Supplement.	f	0	3	8	26	3	40
	%	0.0	7.5	20.0	65.0	7.5	100
10. Item No. 86: Test takers are provided with contemporary self-assessment tools such as European Language Portfolio (ELP).	f	0	1	8	22	9	40
	%	0.0	2.5	20.0	55.0	22.5	100
11. Item No. 58: Test item writers are trained before test administration.	f	2	0	14	14	10	40
	%	5.0	0.0	35.0	35.0	25.0	100

The overall results above showed that 65% (N= 26) of the participants confirmed that the equivalence between different versions of the tests (e.g. year by year) were verified by the institutions at which they were working. Hence, it could be indicated that more than half of the participants were of similar opinion. However, 32.5% (N= 13) of the participants were still not sure whether the private institutions they were working at handled any procedure on verification based upon a pre-defined timely basis, or they were not informed to be so. Not to mention, 2.5% (N= 1) of the participants dissented to the verification of the different versions of the tests, though. Besides, the overall results above showed that 67.5% (N= 27) of the participants confirmed that there were some actions taken after the implementation of each test in order to enhance the quality of teaching and learning by the institutions at which they were working. Hence, it could be indicated that more than half of the participants were of similar opinion. However, 32.5% (N= 13) of the participants were still not sure whether the private institutions they were working at took any actions to improve the quality of teaching and learning, or they were not informed to be so. Furthermore, the results above showed that 65% (N= 26) of the participants confirmed that the institutions at which they were working conducted piloting before administering tests to the target population. Hence, it could be indicated that nearly three out of four of the participants were of similar opinion. Nevertheless, nearly one-third of the participants (N= 11; P= 27.5%) were still not sure whether the private institutions they were working at conducted piloting before administering tests to the target population, or they were not informed to do so.

Additionally, the results above showed that 57.5% (N= 23) of the participants confirmed that the institutions at which they were working used automated scoring machines in marking and grading. Hence, it could be indicated that more than half

of the participants were of similar opinion. Nevertheless, nearly one-fourth of the participants (N= 9; P= 22.5%) were still not sure whether the private institutions they were working at used automated scoring machines in marking and grading, or they were not informed to do so. Otherwise, 20% (N= 8) of the participants claimed that automated scoring machines were not used in marking and grading, though. The results showed that the participants were of different opinion as 42.5% (N= 17) of them were either not sure or disagreed the idea that the private institutions they were working at exploited automated scoring machines. Correlatively, the overall results above showed that 70% (N= 28) of the participants confirmed the use of human scoring after administering tests to the target population. Hence, it could be indicated that nearly three out of four of the participants were of similar opinion. Nevertheless, nearly one-third of the participants (N= 11; P= 27.5%) were still not sure whether the private institutions they were working at used human scoring after administering tests to the target population, or they were not informed to do so. Otherwise, 2.5% (N= 1) of the participants claimed that human scoring was not used after administering tests to the target population, though.

Moreover, the results above showed that 77.5% (N= 31) of the participants confirmed that the institutions at which they were working kept pace with the changes in the current ELT curriculum while designing test items. Hence, it could be indicated that slightly higher than the three-fourth of the participants were of similar opinion. Nevertheless, 17.5% (N= 7) of the participants were still not sure whether the private institutions they were working at kept pace with the changes in the current ELT curriculum while designing test items, or they were not informed to be so. Not to mention, 5% (N= 2) of the participants dissented to keeping pace with the changes in the current ELT curriculum while designing test items, though. Concomitantly, the results above showed that 75% (N= 30) of the participants confirmed that the institutions at which they were working were still in favor of traditional assessment practices. Hence, it could be indicated that nearly three-fourth of the participants were of similar opinion. Besides, 20% (N= 8) of the participants were still not sure whether the private institutions they were working at used traditional assessment practices, or they were not informed to do so. Not to mention, 5% (N= 2) of the participants dissented to the use of traditional assessment practices, though.

On the other hand, the results below showed that 87.5% (N= 35) of the participants confirmed that the institutions at which they were working had a publicly available report on the linking process between the tests in use and the Reference Supplement. Hence, it could be indicated that slightly above than the three-fourth of the participants were of similar opinion. Besides, 7.5% (N= 3) of the participants were still not sure whether the private institutions they were working at had a publicly available report on the linking process between the tests in use and Reference Supplement, or they were not informed to have so. Not to mention, 5% (N= 2) of the participants claimed that the institutions they were working at did not have a publicly available report on the linking process between the tests in use and Reference Supplement, though. Relatively, the results above showed that 72.5% (N= 29) of the participants confirmed that the institutions at which they were working were conducting procedures recommended in the Manual and Reference Supplement as a part of the linkage to the CEFR. Hence, it could be indicated that nearly three-fourth of the participants were of similar opinion. Besides, 20% (N= 8) of the participants were still not sure whether the private institutions they were working at were conducting procedures recommended in the Manual and Reference Supplement as a part of the linkage to the CEFR, or they were not informed to be so. Not to mention, 7.5% (N= 3) of the participants dissented to the fact that the private institutions they were working at were conducting any procedures recommended in the Manual and Reference Supplement as a part of the linkage to the CEFR, though.

As a part of the linkage to the CEFR, it was also asked whether the private institutions enrolled within this study were in favor of using self-assessment tools such as the ELP. Accordingly, the overall results above showed that 77.5% (N= 31) of the participants confirmed that the institutions at which they were working provided their test takers with contemporary self-assessment tools. Hence, it could be indicated that slightly more than three-fourth of the participants were of similar opinion. Besides, 20% (N= 8) of the participants were still not sure whether the private institutions they were working at used contemporary self-assessment tools such as the ELP, or they were not informed to do so. Not to mention, 2.5% (N= 1) of the participants dissented to the use of contemporary self-assessment tools within the private institutions they were working at, though. Last but not least, the

participants were asked whether the private institutions they were working at provided training for their test item writers before administering the tests. In this vein, the overall results above showed that 60% (N= 24) of the participants confirmed that the institutions at which they were working provided their test item writers with training before test administration. Hence, it could be indicated that slightly more than the three-fourth of the participants were of similar opinion. Besides, 35% (N= 14) of the participants were still not sure whether the private institutions they were working at used contemporary self-assessment tools such as the ELP, or they were not informed to do so. Not to mention, 5% (N= 2) of the participants dissented to the use of contemporary self-assessment tools within the private institutions they were working at, though.

As above mentioned, each of the private institutions was also checked separately to detect any implementational difference amidst. Accordingly, the results of each private institution were given below within tables embodied the estimations regarding the exploitation of the EALTA Guidelines by selected private institutions. For each item, frequencies and percentages were given within tables. The results were reported singly, and each item was elaborated in detail, embedding into sub-groups previously defined.

The implementation of the EALTA guidelines by the institution A. An overall estimation regarding the results gained from all of the private institutions were checked and reported together and separately. With this in mind, the overall results of the implementation of the EALTA Guidelines are presented below regarding the case in private institution A.

With a view to quality control and test analyses ascertained by the EALTA, the verification of the equivalence between different versions of the tests (e.g. year by year) was checked initially. Concerning the results of the institution A, it was reported that 36.4% (N= 4) of the participants confirmed the verification of the different versions of the tests. On the other hand, 54.5% (N= 6) of the participants was not sure whether there was any verification process followed by the institution A. Last but not least, 9.1% (N= 1) of the participants dissented to the verification of the different versions of the tests, though. With a mean score of 4.00/ 5.00 (SD= .44), the participants from the institution A held different opinions from each other. Correlatively, it was checked secondarily whether any actions to improve the quality

of teaching and learning were taken after the implementation of each test. Concerning the results of the institution A, it was indicated that 63.6% (N= 7) of the participants confirmed the actions taken to improve the quality of teaching and learning, meaning that more than half of the participants either agreed or strongly agreed upon the moves taken towards the enhancement of the quality of on-going teaching and learning practices. On the other hand, 36.4% (N= 4) of the participants were not sure whether there were any actions taken to improve the quality of teaching and learning after the implementation of each test, indicating that less than half of the participants were not informed about the actions aforementioned. With a mean score of 3.81/ 5.00 (SD= .75), the participants from the institution A were predominantly not sure of the actions aforementioned.

With a view to quality control and test analyses ascertained by the EALTA, it was checked whether the tests were piloted before they were administered to the pre-defined target population. Concerning the results of the institution A, it was yielded that 63.6% (N= 7) of the participants confirmed piloting before administering tests to the target population. On the other hand, 36.4% (N= 4) of the participants was not sure whether the private institution they were working at (institution A) conducted piloting before administering tests to the target population. With a mean score of 4.00/ 5.00 (SD= .89), the participants from the institution A held different opinions from each other. In the same vein, it was checked whether the tests were scored via automated scoring machines after conducting tests. Concerning the results of the institution A, it was yielded that 54.5% (N= 6) of the participants confirmed the use of automated scoring machines in scoring. However, 27.3% (N= 3) dissented to the fact that automated scoring machines were in use at the selected private institutions. One more to note, 18.2% (N= 2) of the participants was not sure whether the private institution they were working at (institution A) used automated scoring machines in marking and grading after administering tests to the target population. It could be stipulated that nearly half of the participants (N= 5; P= 45.5%) were either not sure, or disagreed the fact that automated scoring machines were in use. With a mean score of 3.18/ 5.00 (SD= 1.33), the participants from the institution A held different opinions from each other. Similarly, it was checked whether the tests were scored via human scoring. Concerning the results of the institution A, it was concluded that 81.8% (N= 9) of the participants confirmed the

use of human scoring after administering tests to the target population. On the other hand, 18.2% (N= 2) of the participants was not sure whether the private institution they were working at (institution A) used human scoring after administering tests to the target population. With a mean score of 4.18/ 5.00 (SD= .75), the participants from the institution A held similar opinion with each other.

With a view to review and washback ascertained by the EALTA, it was checked initially whether the test items kept pace with changes in the current ELT curriculum. Concerning the results of the institution A, it was estimated that 63.6% (N= 7) of the participants confirmed keeping pace with the changes in the current ELT curriculum in terms of designing test items. On the other hand, 36.4% (N= 4) of the participants was not sure whether the private institution they were working at (institution A) was meticulous about keeping pace with changes in the current ELT curriculum while designing test items. With a mean score of 3.72/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other. Similarly, as a controlling item for the test items' keeping pace with the changes in the current ELT curriculum, it was checked whether traditional assessment practices were still in use for test takers. In the light of this, the results of the institution A yielded that 54.5% (N= 6) of the participants confirmed that the institutions at which they were working were still in favor of traditional assessment practices. However, 27.3% (N= 3) of the participants were not sure whether the private institution they were working at (institution A) was using traditional assessment practices within. Besides, 18.2% (N= 2) of the participants dissented to the fact that the institution A was still in favor of traditional assessment practices. With a mean score of 4.07/ 5.00 (SD= .81), the participants from the institution A held different opinions from each other.

With a view to linkage to the CEFR ascertained by the EALTA, it was checked whether the tests in use were compatible with the Framework through a publicly available report on the linking process initially. In the light of this, the results of the institution A yielded that 81.8% (N= 9) of the participants confirmed that they had a publicly available report on the linking process between the tests in use and Reference Supplement. However, 18.2% (N= 2) of the participants was not sure whether the private institution they were working at (institution A) had a publicly available report on the linking process between the tests in use and Reference Supplement. With a mean score of 3.82/ 5.00 (SD= .40), the participants from the

institution A held similar opinion with each other. In a similar case, it was checked afterwards whether the tests corresponded to the procedures recommended in the Manual and Reference Supplement. In the light of this, the results of the institution A yielded that 54.5% (N= 6) of the participants was not sure whether the institutions at which they were working were using tests which were compatible with the Manual and Reference Supplement purported by the Framework. However, 27.3% (N= 3) of the participants confirmed that the private institution they were working at (institution A) was using such kind of tests for testing and assessment. Besides, 18.2% (N= 2) of the participants dissented to the fact that the institution A was using tests in tune with the Manual and Reference Supplement. With a mean score of 3.09/ 5.00 (SD= .51), more than half of the participants from the institution A held similar opinion with each other. In the same vein, it was also checked whether test takers were provided with contemporary self-assessment tools such as the ELP. In the light of this, the results of the institution A yielded that 63.6% (N= 7) of the participants confirmed that the institutions at which they were working were using contemporary self-assessment tools such as the ELP. However, 27.3% (N= 3) of the participants were not sure whether the private institution they were working at (institution A) was using contemporary self-assessment tools such as the ELP. Besides, 9.1% (N= 1) of the participants dissented to the fact that the institution A was not using contemporary self-assessment tools such as the ELP. With a mean score of 3.82/ 5.00 (SD= .98), the participants from the institution A held similar opinion with each other.

In relation with the test design and item writing purported by the EALTA, it was checked whether test item writers were trained before test administration. In the light of this, the results of the institution A yielded that 54.5% (N= 6) of the participants confirmed that the institutions at which they were working were training their test item writers before administering the tests. However, 45.5% (N= 5) of the participants was not sure whether the private institution they were working at (institution A) was providing training for their test item writers before test administration. With a mean score of 3.91/ 5.00 (SD= .94), the participants from the institution A held different opinions from each other.

The implementation of the EALTA guidelines by the institution B. An overall estimation regarding the results gained from all of the private institutions

were checked and reported together and separately. With this in mind, the overall results of the implementation of the EALTA Guidelines are listed below regarding the case in private institution B.

With a view to quality control and test analyses ascertained by the EALTA, the verification of the equivalence between different versions of the tests (e.g. year by year) was checked initially. Accordingly, the results of the institution B were reported that 79% (N= 15) of the participants confirmed the verification of the different versions of the tests. On the other hand, 21% (N= 4) of the participants was not sure whether there was any verification process followed by the institution B. With a mean score of 4.16/ 5.00 (SD= .74), the participants from the institution B held similar opinion with each other. Similarly, it was checked secondarily whether any actions to improve the quality of teaching and learning were taken after the implementation of each test. In this vein, it was reported by the results of the institution B that 84.2% (N= 16) of the participants confirmed the actions taken to improve the quality of teaching and learning after the implementation of each test. On the other hand, 15.8% (N= 3) of the participants was not sure whether there was any action taken by the institution B. With a mean score of 4.26/ 5.00 (SD= .73), the participants from the institution B held similar opinion with each other. In the same vein, it was checked whether the tests were piloted before they were administered to the pre-defined target population. Accordingly, the results of the institution B were reported that 73.7% (N= 14) of the participants confirmed piloting before administering tests to the target population. On the other hand, 26.3% (N= 5) of the participants was not sure whether piloting before administering tests to the target population was conducted by the institution B. With a mean score of 4.00/ 5.00 (SD= .75), the participants from the institution B held similar opinion with each other. Correlatively, it was checked whether the tests were scored via automated scoring machines after conducting tests. Accordingly, the results of the institution B were reported that 57.9% (N= 11) of the participants confirmed the use of automated scoring machines in scoring. However, 31.6% (N= 6) of the participants was not sure whether automated scoring machines were used after administering tests to the target population by the institution B. Besides, 10.5% (N= 2) of the participants dissented to the exploitation of automated scoring machines in scoring, though. With a mean score of 3.58/ 5.00 (SD= .84), the participants from the institution B held

different opinions from each other. To note more, it was also checked whether the tests were scored via human scoring. The results of the institution B were reported that 63.2% (N= 12) of the participants confirmed the use of human scoring after administering tests to the target population. On the other hand, 31.6% (N= 6) of the participants was not sure whether human scoring after administering tests to the target population was conducted by the institution B. Additionally, 5.3% (N= 1) of the participants claimed that human scoring was not used in the institution B. With a mean score of 3.79/ 5.00 (SD= .85), the participants from the institution B held similar opinion with each other.

With a view to review and washback ascertained by the EALTA, it was checked initially whether the test items kept pace with changes in the current ELT curriculum. The results of the institution B were reported that all (N= 19; P= 100%) of the participants confirmed keeping pace with the changes in the current ELT curriculum while designing test items. With a mean score of 4.16/ 5.00 (SD= .58), the participants from the institution B held the same opinion with each other. Similarly, as a controlling item for the test items' keeping pace with the changes in the current ELT curriculum, it was also checked whether traditional assessment practices were still in use for test takers. In this respect, the results of the institution B were reported that 68.4% (N= 13) of the participants confirmed using traditional assessment practices within. Additionally, 21.1% (N= 4) of the participants were not sure whether the institution they were working at was in favor of traditional assessment practices, or they were not informed to be so. Not to mention, 20.5% (N= 2) of the participants disagreed the fact that institution B was in favor of traditional assessment practices. With a mean score of 3.58/ 5.00 (SD= .57), the participants from the institution B held similar opinion with each other.

With a view to linkage to the CEFR ascertained by the EALTA, it was checked whether the tests in use were compatible with the Framework through a publicly available report on the linking process. Accordingly, the results of the institution B were reported that 94.7% (N= 18) of the participants confirmed having a publicly available report on the linking process between the tests in use and Reference Supplement. However, 5.3% (N= 1) of the participants were not sure whether the institution they were working at had a publicly available report on the linking process between the tests in use and Reference Supplement, or they were not informed to

have so. With a mean score of 4.11/ 5.00 (SD= .46), the participants from the institution B held similar opinion with each other. In the same vein, it was checked afterwards whether the tests corresponded to the procedures recommended in the Manual and Reference Supplement. Accordingly, the results of the institution B were reported that 100% (N= 19) of the participants confirmed corresponding tests to the procedures recommended in the Manual and Reference Supplement. With a mean score of 4.16/ 5.00 (SD= .42), the participants from the institution B held the same opinion with each other. One more to note, it was also checked whether test takers were provided with contemporary self-assessment tools such as the ELP. Accordingly, the results of the institution B were reported that 89.5% (N= 17) of the participants confirmed providing test takers with contemporary self-assessment tools such as the ELP. However, 10.5% (N= 2) of the participants were not sure whether the institution they were working at provided their test takers with contemporary self-assessment tools, or they were not informed to be so. With a mean score of 4.05/ 5.00 (SD= .52), the participants from the institution B held similar opinion with each other.

In relation with the test design and item writing purported by the EALTA, it was also checked whether test item writers were trained before test administration. Accordingly, the results of the institution B were reported that 63.2% (N= 12) of the participants confirmed that the private institution B provided training for their test item writers before test administration. However, 36.8% (N= 7) of the participants were not sure whether the institution they were working at provided their test item writers with training before administering the tests, or they were not informed to be so. With a mean score of 3.84/ 5.00 (SD= .76), the participants from the institution B held similar opinion with each other.

The implementation of the EALTA guidelines by the institution C. An overall estimation regarding the results gained from all of the private institutions were checked and reported together and separately. With this in mind, the overall results of the implementation of the EALTA Guidelines are given below regarding the case in private institution C.

With a view to quality control and test analyses ascertained by the EALTA, the verification of the equivalence between different versions of the tests (e.g. year by year) was checked initially. Accordingly, the results of the institution C were

reported that 70% (N= 7) of the participants confirmed the verification of the different versions of the tests. On the other hand, 30% (N= 3) of the participants was not sure whether there was any verification process followed by the institution C. With a mean score of 3.70/ 5.00 (SD= .63), the participants from the institution C held similar opinion with each other. Similarly, it was checked secondarily whether any actions to improve the quality of teaching and learning were taken after the implementation of each test. In this vein, the results of the institution C were reported that 40% (N= 4) of the participants confirmed the actions taken to improve the quality of teaching and learning. On the other hand, 60% (N= 6) of the participants was not sure whether there was any action taken by the institution C. With a mean score of 3.50/ 5.00 (SD= .48), the participants from the institution C held different opinions from each other. In the same vein, it was checked whether the tests were piloted before they were administered to the pre-defined target population. Accordingly, the results of the institution C were reported that 50% (N= 5) of the participants confirmed that piloting before administering tests to the target population was conducted by the institution C. On the other hand, 30% (N= 3) of the participants disagreed the idea that piloting before administering tests to the target population was conducted by the institution C. Likewise, 20% (N= 2) of the participants was not sure whether piloting before administering tests. With a mean score of 3.20/ 5.00 (SD= .67), the participants from the institution C held different opinions from each other.

With a view to quality control and test analyses ascertained by the EALTA, it was also checked whether the tests were scored via automated scoring machines after conducting tests. Accordingly, the results of the institution C were reported that 80% (N= 8) of the participants confirmed the use of automated scoring machines by the institution C after administering tests to the target population. On the other hand, 10% (N= 1) of the participants disagreed the fact that automated scoring machines were used by the institution C after administering tests to the target population. Likewise, 10% (N= 1) of the participants was not sure whether automated scoring machines were in use by the institution C. With a mean score of 4.10/ 5.00 (SD= 1.48), the participants from the institution C held similar opinion with each other. In a similar vein, it was checked whether the tests were scored via human scoring. The results of the institution C were reported that 70% (N= 7) of the participants

confirmed that human scoring after administering tests to the target population was conducted by the institution C. On the other hand, 30% (N= 3) of the participants was not sure whether human scoring after administering tests to the target population was conducted by the institution C. With a mean score of 3.90/ 5.00 (SD= 1.49), the participants from the institution C held similar opinion with each other.

With a view to review and washback ascertained by the EALTA, it was checked initially whether the test items kept pace with changes in the current ELT curriculum. Herein, the results of the institution C were reported that 50% (N= 5) of the participants confirmed keeping pace with the changes in the current ELT curriculum while designing test items. On the other hand, 30% (N= 3) of the participants was not sure whether the changes in the current ELT curriculum were followed by the institution C. Besides, 20% (N= 2) of the participants dissented to the fact that their institution kept pace with the changes in the current ELT curriculum. With a mean score of 3.30/ 5.00 (SD= .71), the participants from the institution C held different opinions from each other. Similarly, as a controlling item for the test items' keeping pace with the changes in the current ELT curriculum, it was checked whether traditional assessment practices were still in use for test takers. In this respect, the results of the institution C were reported that 90% (N= 9) of the participants confirmed using traditional assessment practices. On the other hand, 10% (N= 1) of the participants was not sure whether traditional assessment practices were used by the institution C. With a mean score of 3.90/ 5.00 (SD= .82), the participants from the institution C held very similar opinion with each other.

With a view to linkage to the CEFR ascertained by the EALTA, it was checked whether the tests in use were compatible with the Framework through a publicly available report on the linking process initially. Accordingly, the results of the institution C were reported that 80% (N= 8) of the participants confirmed having a publicly available report on the linking process between the tests in use and the Reference Supplement. On the other hand, 20% (N= 2) of the participants dissented to the fact that they had a publicly available report on the linking process between the tests in use and the Reference Supplement within the institution C. With a mean score of 3.60/ 5.00 (SD= .84), the participants from the institution C held similar opinion with each other. In the same vein, it was checked afterwards whether the tests corresponded to the procedures recommended in the Manual and Reference

Supplement. Accordingly, the results of the institution C were reported that 70% (N= 7) of the participants confirmed that the tests corresponded to the procedures recommended in the Manual and Reference Supplement. On the other hand, 20% (N= 2) of the participants was not sure whether the tests in use by the institution C corresponded to the procedures recommended in the Manual and Reference Supplement. Additionally, 10% (N= 1) of the participants dissented to the use of tests within the institution C, which were compatible with the procedures recommended in the Manual and Reference Supplement. With a mean score of 3.60/ 5.00 (SD= .94), the participants from the institution C held similar opinion with each other. One more to note, it was also checked whether test takers were provided with contemporary self-assessment tools such as the ELP. Accordingly, the results of the institution C were reported that 70% (N= 7) of the participants confirmed providing test takers with contemporary self-assessment tools such as the ELP. However, 30% (N= 3) of the participants were not sure whether the institution they were working at provided their test takers with contemporary self-assessment tools such as the ELP, or they were not informed to be so. With a mean score of 4.00/ 5.00 (SD= .74), the participants from the institution C held similar opinion with each other.

In relation with the test design and item writing purported by the EALTA, it was also checked whether test item writers were trained before test administration. Accordingly, the results of the institution C were reported that 60% (N= 6) of the participants confirmed that the private institution C provided training for their test item writers before test administration. However, 20% (N= 2) of the participants were not sure whether the institution they were working at provided their test item writers with training before administering the tests, or they were not informed to be so. Besides, 20% (N= 2) of the participants dissented to the training of test item writers before test administration by the institution C. With a mean score of 3.40/ 5.00 (SD= .57), the participants from the institution C held alike opinions with each other.

The overall picture of the implementation of the EALTA guidelines in language testing and assessment by selected private institutions. As previously mentioned, the EALTA guidelines were summed up in four basic components within the questionnaire used for this study. These components were quality control and test analyses, review and washback, linkage to the CEFR, and

test design and item writing. Composed of 11 test items in total, these four subsections were analyzed separately, and the estimations gained were reported singly. Accordingly, although the number of participants (N= 26; P= 65%) who confirmed that the equivalence between different versions of the tests were verified (e.g. year by year), were higher than those who were not sure whether there was any verification procedure followed by the private institutions previously selected (N= 13; P= 32.5%), the amount of the latter could not be underestimated as there were 40 participants in total. Therefore, it could be stipulated that not all of the participants of this study, the English language teachers who were also working as test (-item) developers at those private institutions, were well aware of the on-going implementations conducted within the institutions.

Additionally, although the number of participants (N= 27; P= 67.5%) who confirmed that there were actions taken in order to enhance the quality of teaching and learning after the implementation of each test, were higher than those who were not sure whether there was any action taken by the private institutions previously selected (N= 13; P= 32.5%), the amount of the latter could not be underestimated as there were 40 participants in total. Therefore, it could be stipulated that not all of the participants of this study, the English language teachers who were also working as test (-item) developers at those private institutions, were well aware of the on-going implementations conducted within the institutions. Besides, although the number of participants (N= 26; P= 65%) who confirmed that tests were piloted before administering them to the target population were higher than those who were not sure about it (N= 11; P= 27.5%) and those who disagreed (N= 3; P= 7.5%), the amount of the following could not be ignored as there were 40 participants in total. Therefore, it could be stipulated that not all of the participants of this study, the English language teachers who were also working as test (-item) developers at those private institutions, were well aware of any implementations regarding piloting conducted within their institutions.

Relatively, although the number of participants (N= 23; P= 57.5%) who confirmed that automated scoring machines were in use were higher than those who were not sure about it (N= 9; P= 22.5%) and those who disagreed (N= 8; P= 20%), the amount of the following could not be underestimated as there were 40 participants in total. Therefore, it could be stipulated that not all of the participants

of this study, the English language teachers who were also working as test (-item) developers at those private institutions, were well aware of the fact that automated scoring machines were in use for the scoring of the results. On the other hand, the number of participants (N= 28; P= 70%) who confirmed the use of human scoring were relatively higher than those who were not sure about it (N= 11; P= 27.5%), or disagreed it (N= 1; P= 2.5%). Therefore, it could be stipulated that a clear majority of the participants stated that human scoring was used for marking and grading after administering the tests to the target population.

Moreover, although the number of participants (N= 31; P= 77.5%) who confirmed that they were keeping pace with the changes in the current ELT curriculum while designing tests and developing test items were higher than those who were not sure about it (N= 7; P= 17.5%), and those who disagreed (N= 2; P= 5%), the amount of the following made us infer that not all of the participants enrolled for this study kept pace with the changes in the current ELT curriculum while designing new test, or developing new test items. Correlatively, the number of participants (N= 30; P= 75%) who confirmed that they were using traditional assessment practices were higher than those who were not sure about it (N= 8; P= 20%), and those who disagreed (N= 2; P= 5%). The amount of the following made us infer that not all of the participants enrolled for this study kept pace with the changes in the current ELT curriculum while designing a new test, or developing new test items as certified with the results gained by previous item. Uninterestingly, traditional assessment methods were still in use by selected private institutions in their testing and assessment practices.

As a part of the linkage to the CEFR, the number of participants (N= 35; P= 87.5%) who confirmed that there was a publicly available report on the linking process between the tests in use and Reference Supplement were higher than those who were not sure about it (N= 3; P= 7.5%), and those who disagreed (N= 2; P= 5%). The amount of the following made us infer that a majority of the participants enrolled for this study affirmed that they had a publicly available report on the linking process between the tests in use and Reference Supplement, such as the CEFR. Similarly, the number of participants (N= 29; P= 72.5%) who confirmed that the tests in use were in harmony with the procedures recommended in the Manual and Reference Supplement were higher than those who were not sure about it (N= 8;

P= 20%), and those who disagreed (N= 3; P= 7.5%). The amount of the following made us infer that a majority of the participants enrolled for this study affirmed that the tests in use by the selected private institutions did correspond to the procedures recommended in the Manual and Reference Supplement of the Framework. In relation to this, the number of participants (N= 31; P= 77.5%) who confirmed that the test takers were provided with contemporary self-assessment tools such as the ELP were higher than those who were not sure about it (N= 8; P= 20%), and those who disagreed (N= 1; P= 2.5%). The amount of the following made us infer that a majority of the participants enrolled for this study affirmed that the selected private institutions were using contemporary self-assessment tools such as the ELP, as well.

Last but not least, although the number of participants (N= 24; P= 60%) who confirmed that the test item writers were trained before test administration were higher than those who were not sure about it (N= 14; P= 35%), and those who disagreed (N= 2; P= 5%), the amount of the following could not be underestimated as there were 40 participants in total. Therefore, it could be stipulated that not all of the participants of this study, the English language teachers who were also working as test (-item) developers at those private institutions, were well aware of the fact that test item writers were trained before test administration.

Do the testing and assessment practices of non-formal English language schools in Turkey correspond to the standards set by the ALTE?

Representing the testing of languages by means of world-leading assessment bodies with a great number of both individual and institutional affiliates, the ALTE has set standards in order to maintain diversity across Europe. Supporting institutions which are producing exams and/or certificates for language learners beyond Europe, the ALTE has established a cycle of common standards starting from the test development, and is followed by task design and item writing, test administration, marking and grading, reporting of the results, test analysis and reporting of the findings respectively. Taking these into consideration, the ALTE Code of Practice and Minimum Standards, as the canons of good practice in language testing and assessment, were the main tenets to define the on-going testing and assessment practices of selected private institutions within the scopes of (a) test construction; (b) administration and logistics; (c) marking and grading; (d)

test analysis; (e) communication with stakeholders; (e) test production; and (f) item writing.

To note beforehand, the items (N= 43) in the questionnaire regarding the 'ALTE Code of Practice' (ALTE, 2010; ALTE, 2017) were numbered as 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, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, and 43. The items were categorized into 7 groups. These groups were named by the ALTE itself as test construction, administration and logistics, marking and grading, test analysis, communication with stakeholders, test production, and item writing. The items in the questionnaire could be listed as given below:

Table 10

Questionnaire Items by the ALTE Code of Practice

Section(s)	Sub-Section(s)	Item(s)
The ALTE Code of Practice	1. Test Construction (10 items)	<p>Item No. 1: The tests are based on a theoretical construct or a model (e.g. communicative competence).</p> <p>Item No. 2: The purpose, context of use, and target population for the tests are appropriately stated.</p> <p>Item No. 3: The tests cover the full range of knowledge and skills relevant and useful to real world situations and authentic language use.</p> <p>Item No. 4: The test scores correlate with a recognized external criterion which measures the same area of knowledge or ability (e.g. the CEFR).</p> <p>Item No. 5: Criteria for selection and training of test constructors and expert judgment are involved both in test construction, and in the review and revision of the tests.</p> <p>Item No. 6: The tests are comparable with parallel examinations across different administrations in terms of content, consistency and grade boundaries.</p> <p>Item No. 7: Evidence of the tests' linkage to an external reference system (e.g. the CEFR) is available through alignment chart.</p> <p>Item No. 8: The purpose of the tests is clearly defined.</p> <p>Item No. 9: The content of the tests is consistent with the stated goal or which the test is being administered.</p> <p>Item No. 10: Discriminant validity sub-scores are supported by means of logical and empirical evidences.</p>
	2. Administration and Logistics (6 items)	<p>Item No. 24: It costs a lot to procure and administer the tests.</p>

2. Administration
and Logistics
(6 items)

Item No. 29: All centers are selected to administer the tests according to clear, transparent, established procedures, and have access to regulations about how to do so.

Item No. 30: Examination papers are delivered in excellent condition, and by secure means to the scoring centers.

Item No. 31: The examination administration system has appropriate support systems (e.g. phone hotline, web services etc.)

Item No. 32: The results are adequately protected by the security, and confidentiality of the results and certificates is enabled.

Item No. 33: The examination system provides support for candidates with special needs.

3. Marking and
Grading (7 items)

Item No. 22: There is a publicly available report on the linking process between tests in use and the Reference Supplement, such as the CEFR.

Item No. 25: As a part of the linkage to the CEFR, the tests correspond to the procedures recommended in the Manual and Reference Supplement.

Item No. 34: Marking is sufficiently accurate and reliable for purpose and type of the test.

Item No. 35: How marking is carried out is documented and explained through raters' reliability estimates.

Item No. 36: The data is collected on an adequate and representative sample of candidates, and not influenced by factors like L1, country of origin, gender, age and ethnic origin.

Item No. 37: Item-level data (e.g. for computing the difficulty, discrimination, reliability, standard errors of measurement of the examination) is collected from an adequate sample of candidates.

Item No. 43: The marking scheme, rubrics, answer keys and rating scales are readily available.

4. Test Analysis
(10 items)

Item No. 11: The test takers' characteristics are clearly defined.

Item No. 12: The tests are appropriate to the overall abilities of the test-takers.

Item No. 13: The tests have been previously tried out on a sample of persons from the same general population as the target test-takers.

Item No. 14: The test results are reliable enough to make accurate decisions.

Item No. 15: The degree of reliability of the test is demonstrated by numerical data.

Item No. 16: The format of the tests is suitable, and its contextual use is clearly defined.

Item No. 17: The test takers are familiar with the actual test format(s).

Item No. 18: The format and features of the tests can be fairly applied in the real testing situations.

Item No. 19: The tests are relevant to the proposed test population and/or to the test item domain.

Item No. 20: The proposed test population/ content resembles the developmental sample closely.

The ALTE Code of Practice	5. Communication with Stakeholders (3 items)	Item No. 38: The test administration system communicates the test results to candidates, and if required, to examination centers (e.g. schools) promptly and clearly. Item No. 39: The stakeholders are informed on the context, purpose, use of the tests, and the overall reliability of the results appropriately. Item No. 40: Stakeholders are informed about how to interpret and use the test results appropriately.	
	6. Test Production (5 items)	Item No. 21: It is easy to produce equivalent or equated forms of the tests being used. Item No. 23: The tests require great deal of training before they are conducted. Item No. 26: The tests are readily available. Item No. 27: The tests are societally and institutionally acceptable. Item No. 28: The tests are acceptable in the eyes of the teachers, parents and administrators.	
	7. Item Writing (2 items)	Item No. 41: The test takers are supplied with different response items (e.g. short answer, sentence correction, gap filling, multiple choice). Item No. 42: The candidates are provided with non-item based task types (e.g. writing tasks, speaking tasks).	
	TOTAL	7 Sub-sections	43 Items

A sum of 43 items, which were above listed in detail, was taken to frequency analysis through descriptive statistics one by one. To add more, for each item, the participants' answers from 3 institutions were estimated and reported singly.

Accordingly, the first main consideration of the ALTE, namely test construction, was composed of 10 core items (item no. 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10). Test construction involved the purpose of the tests in use, the contexts of use, the population targeted at, range of knowledge and skills covered within the tests in use, the theoretical construct or a model based on by the tests in use, a recognized external criterion the test scores correlated with, evidences of the tests' linkage to an external reference system, and the presence of any logical and empirical evidence to support discriminant validity sub-scores. Secondly, the practices nestling administration and logistics were checked with the help of 6 items (item no. 24, 29, 30, 31, 32 and 33). Beforehand, it was asked whether it costed a lot to procure and administer the test. Following that, the regulations of administering were controlled whether there were clear, transparent and established procedures followed by the selected private institutions. Besides, the examination administration

systems were examined whether they had any support systems functioning appropriately. To note more, confidentiality of the results and certificates together with the support for candidates with special needs were asked if there were any in practice.

Thirdly, it was also delved into whether marking and grading systems functioned effectively by means of 7 items (item no. 22, 25, 34, 35, 36, 37 and 43). Within the compass of these items, it was initially asked how easy to score the tests, report the results and interpret the results. Concomitantly, it was also checked whether marking was conducted on a reliable and accurate basis both in implementation and documentation. Any factors such as gender, L1, ethnic origin, country of origin and the like were controlled in order not to influence the results. Besides, it was probed into whether item-level data were collected from an adequate sample of candidates, as well. Not to mention, it was asked whether rubrics, answer keys and rating scales in use were readily available for marking and grading the results. With a view to test analysis, 10 items (item no. 11, 12, 13, 14, 15, 16, 17, 18, 19 and 20) were taken into statistical analysis in order to check the appropriateness of the tests to the test takers. By means of those items, the format of the tests, the degree of reliability, the features of the tests, the contextual use of the tests, the relevancy of the tests to the proposed test population, the results of piloting if conducted, and equivalent or equated forms of the tests if there was any were checked and reported in detail.

Regarding communication with stakeholders, 3 items (item no. 38, 39 and 40) were considered in order to detect whether there was any system communicating the test results to candidates in a prompt and clear way. Besides, it was also checked whether stakeholders were informed on the context, purpose and use of the tests together with the overall reliability of the test results, and were enlightened about how to interpret and use the results gained. Additionally, test production as a sub-section was composed of 5 items (item no. 21, 23, 26, 27 and 28). To elaborate, it was checked whether the tests required a great deal of training before they were conducted. Besides, it was also asked whether the tests were readily available, and societally, institutionally and administratively acceptable.

Last but not least, item writing was composed of 2 items (item no. 41 and 42). Within, it was asked whether the test takers were supplied with different response

items, and whether there were any non-item based task types within those items such as writing tasks, speaking tasks, etc. Before delving into details, table below given embodied the overall estimations regarding the exploitation of the ALTE Code of Practice by selected private institutions. Means, standard deviations and standard errors of mean were given for each item elaborately.

Table 11

The Exploitation of the ALTE Code of Practice by Selected Private Institutions

Section(s)	Item(s)	N	Mean	Std. Error of Mean	Std. Deviation
1. Test Construction	Item No. 1	40	3.95	.147	.932
	Item No. 2	40	4.00	.160	1.01
	Item No. 3	40	3.93	.140	.888
	Item No. 4	40	4.00	.113	.716
	Item No. 5	40	3.88	.130	.822
	Item No. 6	40	3.80	.139	.882
	Item No. 7	40	3.80	.120	.758
	Item No. 8	40	3.95	.109	.687
	Item No. 9	40	3.83	.138	.874
	Item No. 10	40	3.78	.145	.920
2. Administration and Logistics	Item No. 24	40	2.75	.128	.809
	Item No. 29	40	3.73	.134	.847
	Item No. 30	40	4.08	.083	.526
	Item No. 31	40	3.53	.160	1.01
	Item No. 32	40	3.50	.160	1.01
	Item No. 33	40	3.58	1.33	.844
3. Marking and Grading	Item No. 22	40	4.33	.083	.526
	Item No. 25	40	3.90	.106	.672
	Item No. 34	40	3.63	.146	.925
	Item No. 35	40	3.68	.154	.971
	Item No. 36	40	3.93	.126	.797
	Item No. 37	40	3.80	.096	.608
4. Test Analysis	Item No. 43	40	3.95	.080	.504
	Item No. 11	40	3.75	.159	1.01
	Item No. 12	40	3.80	.153	.966
	Item No. 13	40	3.75	.155	.981
	Item No. 14	40	3.55	.164	1.04
	Item No. 15	40	3.68	.110	.694
	Item No. 16	40	3.95	.156	.986
	Item No. 17	40	4.03	.084	.530
Item No. 18	40	3.83	.129	.813	
Item No. 19	40	3.98	.091	.577	
Item No. 20	40	3.85	.122	.770	

5. Communication with Stakeholders	Item No. 38	40	3.70	.120	.758
	Item No. 39	40	3.90	.933	.591
	Item No. 40	40	3.75	.112	.707
6. Test Production	Item No. 21	40	3.88	.096	.607
	Item No. 23	40	4.10	.147	.928
	Item No. 26	40	3.53	.113	.716
	Item No. 27	40	3.90	.100	.632
	Item No. 28	40	4.00	.113	.716
7. Item Writing	Item No. 41	40	3.68	.140	.888
	Item No. 42	40	4.03	.091	.577
TOTAL	7 Sub-sections/ 43 Items	40			

In the light of these, the highest mean score related to the scope of test construction was the item claiming that the test scores were correlated with a recognized external criterion measuring the same area knowledge or ability such as the CEFR (M= 4.00; SD= .71). Likewise, the participants of this study stated that the context of use, and target population for the tests were also appropriately defined in addition to the purpose of the tests in use (M= 4.00; SD= 1.01). Following these, the tests were stipulated to be based on a theoretical construct or a model, such as the communicative competence (M= 3.95; SD= .93). It was followed by the item asserting that the purpose of the tests was clearly defined with one of the highest mean score of all (M= 3.95; SD= .68). Alike, the tests were claimed to cover the full range of knowledge and skills relevant and useful to real world situations and authentic language use with the mean score of 3.93/ 5.00 (SD= .88).

In relation to the sub-section of test construction, it was concluded that criteria for selection and training of test constructors and expert judgment were involved both in test construction, and in the review and revision of the tests (M= 3.88; SD= .82). To some extent, the content of the tests was consistent with the stated goal for which the test was being administered (M= 3.83; SD= .87). As previously confirmed by the test item claiming that the test scores correlated with a recognized external criterion such as the CEFR, the evidence of the tests' linkage to an external reference system (e.g. the CEFR) was stated to be available through alignment

chart by the participants from the selected private institutions (M= 3.80; SD= .75). Relatively, it was concluded that the tests were comparable with parallel examinations across different administrations in terms of content, consistency and grade boundaries with the mean score of 3.80/ 5.00 (SD= .88). Lastly, it was inferred that discriminant validity sub-scores were supported by means of logical and empirical evidence with the lowest mean score of all regarding test construction (M= 3.78; SD= .92).

With a view to administration and logistics, it was claimed by the participants of this study that the examination papers were delivered in excellent condition, and by secure means to the scoring centers with the highest mean score of all (M= 4.08; SD= .52). It was also noted that all centers were selected to administer the tests according to clear, transparent, established procedures, and had access to regulations about how to do so (M= 3.73; SD= .84). Additionally, procuring and administering the tests were not that much costly for them with the lowest mean score of all (M= 2.75; SD= .80). Besides, it was concluded that the examination system provided support for candidates with special needs (M= 3.58; SD= .84). And, that examination system was stipulated to have appropriate support systems such as phone hotline, web services, etc. with the mean score of 3.53/ 5.00 (SD= 1.01). Correlatively, the results were claimed to be adequately protected by the security, and confidentiality of the results and certificates was enabled by selected private institutions with the lowest mean score of all within the scope of administration and logistics (M= 3.50; SD= 1.01).

The sub-section of marking and grading was checked by seven items which yielded the results that it was easy to score the tests, report the test scores and interpret the results with the highest mean score (M= 4.33; SD= .52) out of seven. It was followed by the item purporting that marking scheme, rubrics, answer keys and rating scales were readily available (M= 3.95; SD= .50). Furthermore, it was stipulated by the participants of this study that the data was collected on an adequate and representative sample of candidates, and not influenced by factors like L1, country of origin, gender, age and ethnic origin with the mean score of 3.93/ 5.00 (SD= .79). Although it was noted by the item of administration and logistics stating that procuring and administering the tests were not that much costly, it was concluded by the item of marking and grading that scoring the tests was costly with

the mean score of 3.90/ 5.00 (SD= .67). Additionally, item-level data (e.g. for computing the difficulty, discrimination, reliability and standards errors of measurement of the examination) were stipulated to be collected from an adequate sample of candidates with the mean score of 3.80/ 5.00 (SD= .60). On the other hand, how marking was carried out was noted to be documented and explained through raters' reliability estimates with the second lowest mean score (M= 3.68; SD= .97) of all. The lowest mean score of the sub-section of marking and grading was estimated by the results that marking was sufficiently accurate and reliable for purpose and type of the test (M= 3.63; SD= .92).

Within the scope of test analysis, it was concluded that the test takers were familiar with the actual test format(s) with the highest mean score out of ten items (M= 4.03; SD= .53). It was followed by the item stipulating that the tests were relevant to the proposed test population and/or to the test item domain with the mean score of 3.98/ 5.00 (SD= .57). The format of the tests was noted to be suitable, and its contextual use was found clear by the participants of this study with the mean score of 3.95/ 5.00 (SD= .98).

Moreover, the format and features of the tests were claimed to be fairly applied in the real testing situations with the mean score of 3.83/ 5.00 (SD= .81). Following that, the results of this study yielded that the tests were found appropriate to the overall abilities of the test takers with the mean score of 3.80/ 5.00 (SD= .96). The results of the sub-section of test analysis supported the idea that the tests were previously tried out on a sample of persons from the same general population as the target test-takers with the mean score of 3.75/ 5.00 (SD= .98). Likewise, it was also concluded that the test takers' characteristics were clearly defined with the same mean score (M= 3.75; SD= 1.01). The second lowest means score was for the item supporting that the degree of reliability of the test was demonstrated by numerical data (M= 3.68; SD= .69). At last, the lowest mean score was noted with the item claiming that the test results were reliable enough to make accurate decisions (M= 3.55; SD= 1.04).

As another sub-section of the ALTE Code of Practice, communication with stakeholders were checked with the help of three items in the questionnaire. Accordingly, it was gained by the results of this study that the stakeholders were stated to be informed on the context, purpose, use of the tests, and the overall

reliability of the test results appropriately with the highest mean score ($M= 3.90$; $SD= .59$) of all. Following that, stakeholders were noted to be informed about how to interpret and use the test results appropriately with the mean score of $3.75/ 5.00$ ($SD= .70$). Lastly, it was also concluded that the test administration system was claimed to communicate the test results to candidates, and if required, to examination centers (e.g. schools) promptly and clearly with the lowest mean score of all ($M= 3.70$; $SD= .75$) regarding the sub-section of communication with stakeholders.

For the sub-section of test production, the highest mean score was estimated as $4.10/ 5.00$ ($SD= .92$), indicating that the tests in use required a great deal of training before they were conducted. It was followed by the second highest mean score of $4.00/ 5.00$ ($SD= .71$), supporting that the tests were acceptable in the eyes of the teachers, parents and administrators. In relation to this, the tests were noted to be societally and institutionally acceptable with the mean score of $3.90/ 5.00$ ($SD= .63$). Besides, it was concluded by the results of this study that it was easy to produce equivalent or equated forms of the tests being used with the mean score of $3.88/ 5.00$ ($SD= .60$). Lastly, the tests in use were noted to be readily available with the lowest mean score ($M= 3.53$; $SD= .71$) of all regarding the sub-section of test production.

Last but not least, for the sub-section of item writing, the highest mean score was estimated as $4.03/ 5.00$ ($SD= .57$), indicating that the candidates were provided with non-item based task types, such as writing tasks, speaking tasks, and the like. On the other hand, the lowest mean score was estimated as $3.68/ 5.00$ ($SD= .88$), supporting that the test takers were supplied with different response items, such as short answer, sentence correction, gap filling and multiple choice to some extent. Therefore, it could be stipulated that although the candidates were provided with non-item based task types, they were not catered with different types of response items.

Keeping these in mind, each sub-section was analyzed separately for each of the selected private institutions. The results were elaborated in detail, and the tables for each were given one by one. At first, an overall estimation regarding the results gained from all of the private institutions were checked and reported together. Following that, the results of each private institution were checked and

reported separately by means of frequencies and percentages given within tables. With these in mind, the table below showed the overall results in a sub-section based order before delving into the results of each private institutions in detail. Each item was reported underneath singly, and the overall estimations were supported by their implications.

Table 12

The Implementation of the ALTE Code of Practice by Selected Private Institutions

The Implementation of the ALTE Code of Practice		Strongly Disagree	Disagree	Not Sure	Agree	Strongly Agree	TOTAL
1. Item No. 1: The tests are based on a theoretical construct or a model (e.g. communicative competence).	f	2	0	6	22	10	40
	%	5.0	0.0	15.0	55.0	25.0	100
2. Item No. 2: The purpose, context of use, and target population for the tests are appropriately stated.	f	0	6	2	18	14	40
	%	0.0	15.0	5.0	45.0	35.0	100
3. Item No. 3: The tests cover the full range of knowledge and skills relevant and useful to real world situations and authentic language use.	f	2	0	5	25	8	40
	%	5.0	0.0	12.5	62.5	20.0	100
4. Item No. 4: The test scores correlate with a recognized external criterion which measures the same area knowledge or ability (e.g. the CEFR).	f	0	0	10	20	10	40
	%	0.0	0.0	25.0	50.0	25.0	100
5. Item No. 5: Criteria for selection and training of test constructors and expert judgment are involved both in test construction, and in the review and revision of the tests.	f	0	2	10	19	9	40
	%	0.0	5.0	25.0	47.5	22.5	100
6. Item No. 6: The tests are comparable with parallel examinations across different administrations in terms of content, consistency and grade boundaries.	f	0	3	11	17	9	40
	%	0.0	7.5	27.5	42.5	22.5	100
7. Item No. 7: Evidence of the tests' linkage to an external reference system (e.g. the CEFR) is available through alignment chart.	f	0	0	16	16	8	40
	%	0.0	0.0	40.0	40.0	20.0	100
8. Item No. 8: The purpose of the tests is clearly defined.	f	0	1	8	23	8	40
	%	0.0	2.5	20.0	57.5	20.0	100
9. Item No. 9: The content of the tests is consistent with the stated goal for which the test is being administered.	f	0	4	7	21	8	40
	%	0.0	10.0	17.5	52.5	20.0	100
10. Item No. 10: Discriminant validity sub-	f	0	5	7	20	8	40

scores are supported by means of logical and empirical evidence.	%	0.0	12.5	17.5	50.0	20.0	100
11. Item No. 24: It costs a lot to procure and administer the tests.	f	1	15	18	5	1	40
	%	2.5	37.5	45.0	12.5	2.5	100
12. Item No. 29: All centers are selected to administer the tests according to clear, transparent, established procedures, and have access to regulations about how to do so.	f	0	3	12	18	7	40
	%	0.0	7.5	30.0	45.0	17.5	100
13. Item No. 30: Examination papers are delivered in excellent condition, and by secure means to the scoring centers.	f	0	0	4	29	7	40
	%	0.0	0.0	10.0	72.5	17.5	100
14. Item No. 31: The examination administration system has appropriate support systems (e.g. phone hotline, web services etc.).	f	2	3	13	16	6	40
	%	5.0	7.5	32.5	40.0	15.0	100
15. Item No. 32: The results are adequately protected by the security, and confidentiality of the results and certificates is enabled.	f	2	4	11	18	5	40
	%	5.0	10.0	27.5	45.0	12.5	100
16. Item No. 33: The examination system provides support for candidates with special needs.	f	0	4	14	17	5	40
	%	0.0	10.0	35.0	42.5	12.5	100
17. Item No. 22: It is easy to score the tests, report the test scores and interpret the results.	f	0	0	1	25	14	40
	%	0.0	0.0	2.5	62.5	35.0	100
18. Item No. 25: It costs a lot to score the tests.	f	2	15	16	7	0	40
	%	5.0	37.5	40.0	17.5	0.0	100
19. Item No. 34: Marking is sufficiently accurate and reliable for purpose and type of the test.	f	0	7	6	22	5	40
	%	0.0	17.5	15.0	55.0	12.5	100
20. Item No. 35: How marking is carried out is documented and explained through raters' reliability estimates.	f	2	0	15	15	8	40
	%	5.0	0.0	37.5	37.5	20.0	100
21. Item No. 36: The data is collected on an adequate and representative sample of candidates, and not influenced by factors like L1, country of origin, gender, age and ethnic origin.	f	0	0	14	15	11	40
	%	0.0	0.0	35.0	37.5	27.5	100
22. Item No. 37: Item-level data (e.g. for computing the difficulty, discrimination, reliability and standards errors of measurement of the examination) is collected from an adequate sample of candidates.	f	0	0	12	24	4	40
	%	0.0	0.0	30.0	60.0	10.0	100
23. Item No. 43: The marking scheme, rubrics, answer keys and rating scales are readily available.	f	0	0	6	30	4	40
	%	0.0	0.0	15.0	75.0	10.0	100

24. Item No. 11: The test takers' characteristics are clearly defined.	f	1	4	8	18	9	40
	%	2.5	10.0	20.0	45.0	22.5	100
25. Item No. 12: The tests are appropriate to the overall abilities of the test takers.	f	0	5	8	17	10	40
	%	0.0	12.5	20.0	42.5	25.0	100
26. Item No. 13: The tests have been previously tried out on a sample of persons from the same general population as the target test-takers.	f	0	5	10	15	10	40
	%	0.0	12.5	25.0	37.5	25.0	100
27. Item No. 14: The test results are reliable enough to make accurate decisions.	f	2	4	10	18	6	40
	%	5.0	10.0	25.0	45.0	15.0	100
28. Item No. 15: The degree of reliability of the test is demonstrated by numerical data.	f	0	3	9	26	2	40
	%	0.0	7.5	22.5	65.0	5.0	100
29. Item No. 16: The format of the tests is suitable, and its contextual use is clearly defined.	f	0	5	5	17	13	40
	%	0.0	12.5	12.5	42.5	32.5	100
30. Item No. 17: The test takers are familiar with the actual test format(s).	f	0	0	6	29	5	40
	%	0.0	0.0	15.0	72.5	12.5	100
31. Item No. 18: The format and features of the tests can be fairly applied in the real testing situations.	f	0	3	8	22	7	40
	%	0.0	7.5	20.0	55.0	17.5	100
32. Item No. 19: The tests are relevant to the proposed test population and/or to the test item domain.	f	0	0	7	27	6	40
	%	0.0	0.0	17.5	67.5	15.0	100
33. Item No. 20: The proposed test population/ content resemble the developmental sample closely.	f	0	1	12	19	8	40
	%	0.0	2.5	30.0	47.5	20.0	100
34. Item No. 38: The test administration system communicates the test results to candidates, and if required, to examination centers (e.g. schools) promptly and clearly.	f	0	2	13	20	5	40
	%	0.0	5.0	32.5	50.0	12.5	100
35. Item No. 39: The stakeholders are informed on the context, purpose, use of the tests, and the overall reliability of the test results appropriately.	f	0	0	9	26	5	40
	%	0.0	0.0	22.5	65.0	12.5	100
36. Item No. 40: Stakeholders are informed about how to interpret and use the test results appropriately.	f	0	1	13	21	5	40
	%	0.0	2.5	32.5	52.5	12.5	100
37. Item No. 21: It is easy to produce equivalent or equated forms of the tests being used.	f	0	0	10	25	5	40
	%	0.0	0.0	25.0	62.5	12.5	100
38. Item No. 23: The tests require a great deal of training before they are conducted.	f	1	2	3	20	14	40
	%	2.5	5.0	7.5	50.0	35.0	100
39. Item No. 26: The tests are readily available.	f	0	1	21	14	4	40
	%	0.0	2.5	52.5	35.0	10.0	100

40. Item No. 27: The tests are societally and institutionally acceptable.	f	0	0	10	24	6	40
	%	0.0	0.0	25.0	60.0	15.0	100
41. Item No. 28: The tests are acceptable in the eyes of teachers, parents and administrators.	f	0	0	10	20	10	40
	%	0.0	0.0	25.0	50.0	25.0	100
42. Item No. 41: The test takers are supplied with different response item (e.g. short answer, sentence correction, gap filling, multiple choice).	f	0	3	15	14	8	40
	%	0.0	7.5	37.5	35.0	20.0	100
43. Item No. 42: The candidates are provided with non-item based task types (e.g. writing tasks, speaking tasks).	f	0	0	6	27	7	40
	%	0.0	0.0	15.0	67.5	17.5	100

The overall results above showed that 80% (N= 32) of the participants confirmed that the tests in use were based on a theoretical construct, or a model. On the other hand, 15% (N= 6) of the participants were still not sure whether the selected private institutions were using tests based on a theoretical construct, or a model. Additionally, 5% (N= 2) of the participants dissented to this fact although it was just a smidgen. Hence, it could be indicated that most of the participants accepted the presence of a theoretical construct, or a model within the tests in use. Above, the tests were claimed to have a theoretical basis by most of the participants.

Correlatively, 80% (N= 32) of them also agreed that the tests in use had a purpose, context of use and target population, which were all appropriately stated. On the other hand, 15% (N= 6) of the participants claimed that the selected private institutions were using tests with no pre-defined purpose, context of use and target population. Additionally, 5% (N= 2) of the participants were not sure of this fact. At the very same, 82.5% (N= 33) of the participants supported that the tests in use covered the full range of knowledge and skills relevant and useful to real world situations and authentic language use. On the other hand, 12.5% (N= 5) of the participants were still not sure whether the selected private institutions were using tests covered the skills aforementioned. Additionally, 5% (N= 2) of the participants dissented to this fact. Hence, it could be indicated that most of the participants accepted the presence of authentic language use and real world situations embedded into tests by means of knowledge and skills covered relevantly.

In terms of correlating the tests scores with a recognized external criterion such as the CEFR, it could be stipulated that most of the participants (N= 30; P=

75%) accepted the presence of an external criterion which was used to measure the same area knowledge or ability. However, 25% (N= 10) of the participants were not sure about the correlation between the tests in use and an external criterion applied. Correlatively, most of the participants (N= 24; P= 60%) agreed that there was an alignment chart as an evidence for the linkage of tests to an external reference system. However, nearly half of them (N= 16; P= 40%) was not sure whether there was an alignment chart or not.

Moreover, 70% (N= 28) of the participants stated that criteria for selection and training of test constructors and expert judgment were involved both in test construction, and in the review and revision of the tests. On the other hand, 25% (N= 10) of them were not sure of it. Additionally, 5% (N= 2) of them disagreed the fact that there were criteria defined for selection and training of test constructors and expert judgment, which were all involved both in test construction, and in the review and revision of the tests. In relation to this, 65% (N= 26) of the participants claimed that the tests were comparable with parallel examinations across different administrations in terms of content, consistency and grade boundaries. However, 27.5% (N= 11) of them were not sure of it. One more to note, 7.5% (N= 3) of them dissented to the presence of comparisons amidst parallel examinations across different administrations.

Besides, the majority of the participants (N= 31; P= 77.5%) stated that the tests in use had a purpose, which was clearly defined. Relatively, 72.5% (N= 29) of the participants confirmed that the content of the tests was consistent with the stated goal for which the test was being administered. Additionally, 70% (N= 28) of the participants claimed that discriminant validity sub-scores were supported by means of logical and empirical evidence. Hence, it could be inferred that the tests in use were marked as valid by means of logical and empirical evidence.

With a view to administration and logistics, 45% (N= 18) of the participants stated that they were not sure whether it costed a lot to procure and administer the tests. On the other hand, 40% (N= 16) of them disagreed that it was costly to procure and administer the tests. Additionally, 15% (N= 6) of them agreed that it was costly to procure and administer the tests. Besides, 62.5% (N= 25) of the participants asserted that all centers were selected to administer the tests according to clear, transparent, established procedures, and have access to regulations about how to

do so. However, 30% (N= 12) of the them were not sure whether the centers selected for test administration had an access to regulations how to do aforementioned implementations. Judicious amount of them (N= 3; P= 7.5%) disagreed with it, though.

Furthermore, a great majority of the participants (N= 36; P= 90%) stated that examination papers were delivered in excellent condition, and by secure means to the scoring centers after the administration of testing process. However, the rest (N= 4; P= 10%) was not sure whether excellent conditions were met through the delivery of examination papers. Interestingly, solely a little more than half of the participants agreed on confidentiality of the results when it came to talking about security. Herein, it was stated by 57.5% (N= 23) of the participants that the results were adequately protected by the security, and confidentiality of the results and certificates was enabled. On the other hand, 32.5% (N= 13) of them were not sure whether the tests were adequately protected by the security in order to keep the confidentiality of the results. Additionally, 12.5% (N= 5) of them disagreed it.

Additionally, 55% (N= 22) of the participants supported that the examination administration system had appropriate support systems, such as phone hotline, web services etc. On the other hand, 32.5% (N= 13) of them was not sure whether there were any support systems aforementioned. Additionally, 12.5% (N= 5) of them dissented to the fact that the private institution they were working at had suitable support systems for test administration. Besides, 55% (N= 22) of the participants affirmed that the examination system in use provided support for candidates with special needs. However, 35% (N= 14) of them were not sure about it. In addition to this, 10% (N= 4) of them dissented that they were catering support for candidates with special needs.

In relation with marking and grading, it was pointed out by nearly all of the participants of this study (N= 39; P= 97.5%) that it was easy to score the tests, report the test scores, and interpret the results. Additionally, 42.5% (N= 17) of them confirmed that it was not costly to score the tests. However, 40% (N= 16) of them was not sure whether it costed a lot to score the tests, or not. Last but not least, 17.5% (N= 7) of them stated that it was costly to score the tests. Relatively, marking was found sufficiently accurate and reliable for purpose and type of the test by the majority of the participants (N= 27; P= 67.5%). Similar results were gathered from

the ones who disagreed (N= 7; P= 17.5%), and those who were not sure (N= 6; P= 15%) about it, though. Supporting this, 57.5% (N= 23) of the participants confirmed that how marking was carried out was documented and explained by means of raters' reliability estimates. However, 37.5% (N= 15) of them were not sure whether the marking process was documented and elaborated through inter-raters' reliability ratings.

Within the scope of data collection, it was marked by the majority of the participants (N= 34; P= 85%) that the marking schemes, rubrics, answer keys and rating scales were readily available. Additionally, 65% (N= 26) of the participants confirmed that the data was collected on an adequate and a representative sample of the whole candidates regardless of any external factor such as country of origin, age, gender, L1 or ethnicity. The rest (N= 14; P= 35%) was not sure whether such kind of external factors that might penetrate into marking and grading were neglected. Correlatively, 70% (N= 28) of the participants confirmed that item-level data was collected from an adequate sample of candidates for the goodness of estimating reliability, item difficulty, discrimination and standard errors of measurement of the tests in use. However, the rest (N= 12; P= 30%) was not sure whether such estimations were calculated after the implementation of each test.

With a view to test analysis, it was confirmed by the majority of the participants (N= 34; P= 85%) that all of the test takers were somewhat familiar with the actual test format(s). Similarly, 82.5% (N= 33) of them confirmed that the tests were relevant to the proposed test population and to the item domain. This was why the test format(s) was found suitable with its clearly-defined contextual use by the 75% (N= 30) of them. The test format(s) was also found transferrable to the real testing situations by the 72.5% (N= 29) of the overall sample. Besides, the tests were found suitable to the overall abilities of the test takers with the majority of the participants (N= 27; P= 67.5%). The same estimated population (N= 27; P= 67.5%) supported that test takers' characteristics were clearly defined before test administration. In this context, 67.5% (N= 27) of the participants confirmed that the tests were appropriate to the overall abilities of the test takers; however, 20% (N= 8) of them was not sure whether the tests in use were suitable in terms of test takers' overall abilities.

Concomitantly, 62.5% (N= 25) of the participants stated that the tests have been previously tried out on a sample of persons from the same general population as the target test-takers. Yet, 25% (N= 10) of them was not sure whether the tests in use were previously put into practice on a similar sample. Additionally, 12.5% (N= 5) of them disagreed with it, though. In the same vein, 67.5% (N= 27) of the participants confirmed that the proposed test population was similar to that of developmental sample but 30% (N= 12) of them was not sure about the similarity between the proposed test population and that of developmental sample. One more to note, 2.5% (N= 1) of them disagreed the idea that there was a close resemblance between the aforementioned sample populations.

Moreover, 70% (N= 28) of the participants confirmed that the degree of test reliability was estimated and demonstrated by numerical data. However, 22.5% (N= 9) was not sure whether reliability estimates were shown numerically. From this point of view, the test results were found reliable enough to fair decisions afterwards by the 60% (N= 24) of the overall sample. Herein, 25% (N= 10) of them was not sure about the reliability of the test scores for the fairness of the ultimate decisions. Additionally, 15% (N= 6) of them disagreed with it, though. Furthermore, 62.5% (N= 25) of the participants asserted that the test administration system communicated the test results to candidates and/or examination centers if needed in a prompt and clear way. However, 32.5% (N= 13) of them was not sure whether the test results were rendered to candidates and/or examination centers via test administration systems.

Besides, 65% (N= 26) of the participants confirmed that the stakeholders were informed about interpreting and using the test results in an appropriate way. On the other hand, 32.5% (N= 13) was not sure whether the stakeholders were instructed on how to use and interpret the test results. One more to note on the issue of communication with stakeholders, 77.5% (N= 31) of the participants stated that the stakeholders were acquainted with the purpose and context of the tests together with the reliability of the results in an appropriate way. Notwithstanding, 22.5% (N= 9) of them were not sure whether the stakeholders were informed about the topics aforementioned.

With respect to test production, it was confirmed by the 85% (N= 34) of the participants that the tests in use required training before administration. Similarly,

75% (N= 30) of the participants confirmed that it was easy to produce tests alike. However, 25% (N= 10) of them were not sure whether tests equivalent to each other were produced by the private institutions they were working at, or not. Moreover, the tests in use were found acceptable both by the society and institutions by the majority of the participants (N= 30; P= 75%). Herein, 25% (N= 10) of them were not sure whether the tests required a great deal of training before conducting. At the very same, the tests were also found acceptable in the eyes of teachers, parents and administrators by the majority of the participants (N= 30; P= 75%). Yet, 25% (N= 10) of them were not sure whether the tests were acceptable in the eyes of the ranks aforementioned. One more to note, 52.5% (N= 21) of the participants was not sure whether the tests in use were readily available. It was laced with the 45% (N= 28) of the participants who confirmed that the tests were readily available. Not to mention, 2.5% (N= 1) of them dissented to the fact that the tests in use were readily available.

In terms of item writing, 85% (N= 34) of the participants confirmed that the candidates were supplied with non-item based task types such as speaking tasks, writing tasks etc. Besides, 15% (N= 6) of the participants were not sure whether non-item based task types were rendered to the candidates. Furthermore, 75% (N= 32) of the participants confirmed that different response items, such as sentence correction, multiple choice, gap filling, short answer and the like were rendered to the test takers in terms of item variety. However, 37.5% (N= 15) of them were not sure whether different types of response items were used in test to enable item variety for the test takers. One more to note, 7.5% (N= 3) of them disagreed with the fact that the test takers were provided with different response items.

As above mentioned, each of the private institutions was also checked separately to detect any implementational difference amidst. Accordingly, the results of each private institution were given below within tables embodied the estimations regarding the exploitation of the ALTE Code of Practice and Minimum Standards by selected private institutions. For each item, frequencies and percentages were given within tables. The results were reported singly, and each item was elaborated in detail, embedding into sub-groups previously defined.

The implementation of the ALTE code of practice by the institution A.

An overall estimation regarding the results gained from all of the private institutions

were checked and reported together and separately. With this in mind, the overall results of the implementation of the ALTE Code of Practice are given below regarding the case in private institution A.

With a view to test construction ascertained by the ALTE, it was initially checked whether the tests were grounded upon any theoretical construct. Concerning the results of the institution A, it was reported that nearly 75% (N= 8) of the participants confirmed the basis of a pre-defined model or a notional construct. On the other hand, 27.3% (N= 3) of the participants was not sure about it. With a mean score of 4.00/ 5.00 (SD= .44), the participants from the institution A held similar opinion with each other. Hence, it could be stipulated that participants from the institution A predominantly accepted the presence of a theoretical construct or a model applied within the tests. Similarly, it was checked secondarily whether test purpose, its contextual use and population targeted at were defined appropriately. Concerning the results of the institution A, it was concluded that 63.7% (N= 7) of the participants confirmed that all aforementioned were defined properly. On the other hand, 18.2% (N= 2) of the participants were not sure whether the purpose together with the context of use of the test were clearly defined besides its target population. At the very same, 18.2% (N= 2) of the participants disagreed nestling to the fact that all aforementioned were not appropriately described. With a mean score of 3.73/ 5.00 (SD= .75), the participants from the institution A held similar opinion with each other. Relatively, 63.6% (N= 7) of the participants confirmed that the test purpose was clearly stated. On the other hand, 36.4% (N= 4) of them was not sure whether the purpose of the tests was given clearly. With a mean score of 3.73/ 5.00 (SD= .89), the participants from the institution A held similar opinion with each other. By the same token, slightly higher than the half of the participants from the institution A (N= 6; P= 54.6%) confirmed that the content of the tests was in correlation with the goal previously stated. However, 27.3% (N= 3) of them was not sure whether there was a correlation between the stated goal and test content. Besides, 18.2% (N= 2) of them dissented that there was a correlation between the stated goal and test content. With a mean score of 3.45/ 5.00 (SD= .89), the participants from the institution A held different opinions from each other.

With a view to test construction ascertained by the ALTE, it was also checked whether the tests encapsulated the full range of knowledge, skills and authentic

language use, which could be applied in real world situations. Concerning the results of the institution A, it was yielded that 72.7% (N= 8) of the participants confirmed it. On the other hand, 27.3% (N= 3) of the participants was not sure whether the private institution they were working at (institution A) prepared tests that covered relevance among knowledge, skills and authenticity in terms of language use. With a mean score of 3.82/ 5.00 (SD= .89), the participants from the institution A held similar opinion with each other. In a very similar vein, it was checked whether the tests in use were coupled with an external reference system, which was herein the Framework. Concerning the results of the institution A, it was yielded that 81.8% (N= 9) of the participants confirmed the relationship between the tests in use and CEFR. However, 18.2% (N= 2) of the participants were not sure whether an external criterion was followed and aligned while preparing the tests in use. With a mean score of 3.91 / 5.00 (SD= 1.33), the participants from the institution A predominantly confirmed the alignment of the tests to the Framework, holding similar opinion with each other. Interestingly, when they were asked about the evidences of the tests' linkage to the CEFR, 63.6% (N= 7) of the participants from the institution A were not sure whether there was an alignment chart available. Merely 26.4% (N= 4) of them confirmed the presence of an alignment chart. With a mean score of 3.36/ 5.00 (SD= .75), the participants from the institution A held similar opinion with each other. In addition to these, it was also asked whether the tests in use were comparable with parallel examinations in recognition of content, consistency and parameters of grading. Herein, 63.6% (N= 7) of the participants confirmed the aforementioned comparability. However, 36.4% (N= 4) of them was not sure whether there was a comparison made amidst examinations parallel to each other within the scope of content, consistency and grading parameters. With a mean score of 3.73/ 5.00 (SD= .75), the participants from the institution A held similar opinion with each other.

With a view to test construction ascertained by the ALTE, it was additionally checked whether there were criteria used for the selection and training of the test constructors, which were involved both in the processes of test construction and revision. In the light of this, 63.6% (N= 6) of the participants from the institution A confirmed that criteria for selection and training of test item writers were involved in the processes aforementioned, and were also laced with expert judgment. However, 36.4% (N= 4) of them was not sure whether there were criteria pre-defined for test

construction and revision of the tests. With a mean score of 3.73/ 5.00 (SD= .75), the participants from the institution A held similar opinion with each other. Moreover, slightly higher than the half of the participants from the institution A (N= 6; P= 54.6%) supported that logical and empirical evidence were testified for the estimations of discriminant validity sub-scores. Yet, 27.3% (N= 3) of them was not sure whether these evidences were given. Additionally, 18.2% (N= 2) of them claimed that there were not any logical and empirical evidence given to support discriminant validity sub-scores. With a mean score of 3.55/ 5.00 (SD= .75), the participants from the institution A held different opinions from each other.

With a view to administration and logistics ascertained by the ALTE, 45.5% (N= 5) of the participants was not sure whether the tests in use were costly to procure and administer, composing the largest proportion of all. Besides, 36.4% (N= 4) of them dissented to the fact that it costed a lot to administer the tests. Additionally, the rest (N= 2; P= 18.2%) confirmed that it was costly to procure and administer the tests. With a mean score of 2.73/ 5.00 (SD= .90), the participants from the institution A held different opinions from each other. Similarly, it was asked whether the test administration centers were selected concerning transparency with established set of procedures and regulations. Herein, 63.7% (N= 7) of the participants confirmed that the aforementioned centers selected accordingly. On the other hand, 27.3% (N= 3) of them was not sure whether these criteria were followed for the selection of test administration centers. Not to mention, 9.1% (N= 1) of them dissented to this fact, though. With a mean score of 3.91/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other. In addition to these, it was also checked whether test administration system had proper support systems. In this vein, slightly higher than the half of the participants (N= 6; P= 54.5%) was not sure whether there were any support systems such as phone hotline, or web services. On the other hand, 36.4% (N= 4) of them confirmed that they had such support systems. One more to note, 9.1% (N= 1) of them stated that there were not any support systems as aforementioned. With a mean score of 3.36/ 5.00 (SD= .65), the participants from the institution A held different opinions from each other. At that point, the participants were also asked whether the test administration system pursued within the private institutions they were working at had support systems for the candidates with special needs. Herein, the same results were

gained from the ones who were not sure about it, and those who confirmed it. To elaborate, 45.5% (N= 5) of the participants was not sure whether the candidates with special needs were supported. However, the same ratio of participants (N= 5; P= 45.5%) confirmed that there were support systems for the candidates with special needs. Besides, 9.1% (N= 1) of the participants disagreed with it. With a mean score of 3.36/ 5.00 (SD= .65), the participants from the institution A held different opinions from each other.

With special interest to administration and logistics ascertained by the ALTE, it was also asked whether the results were protected appropriately by the security in order to keep the confidentiality of the test results and/or certificates. Herein, 63.6% (N= 7) of the participants confirmed that the security system in use could enable the confidentiality of the results. Yet, 27.3% (N= 3) of them were not sure whether the security support system could keep the confidentiality of the test results and/or certificates. One more to note, 9.1% (N= 1) of them were opposed to it, asserting that the results were not adequately protected. With a mean score of 3.64/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other. Correlatively, it was asked whether the tests were delivered to the scoring centers in good conditions thanks to security systems adopted. At that point, the majority of the participants from the institution A (N= 10; P= 90.9%) confirmed that the delivery was done properly and securely to the scoring centers. Not to mention, 9.1% (N= 1) of them was not sure whether security systems in use could make the test results delivered to the scoring centers appropriately. With a mean score of 4.27/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other.

Within the scope of marking and grading as ascertained by the ALTE, it was initially asked whether it was easy to score the tests, report the scores and interpret the results. Herein, all of the participants from the institution A (N= 11; P= 100%) agreed upon it. Therefore, it could be stipulated that scoring the tests and interpreting the results were easy together with reporting the scores. With a mean score of 4.55/ 5.00 (SD= .65), the participants from the institution A held the same opinion with each other. Additionally, it was also asked whether it was costly to score the tests. In this vein, slightly above than the half of the participants (N= 6; P= 54.6%) stated that it was not costly to score the tests. On the other hand, 36.4% (N= 4)

confirmed that scoring the tests was costly. One more note, 9.1% (N= 1) of them was not sure whether it costed too much to score the tests. With a mean score of 2.64/ 5.00 (SD= 1.21), the participants from the institution A held different opinions from each other. In the same vein, it was also asked whether marking was done accurately to yield reliable results in terms of test purpose and type. Herein, 63.7% (N= 7) of the participants confirmed that marking was adequately accurate and reliable. However, 27.3% (N= 3) of the participants dissented to it, stating that marking was not sufficiently reliable for test purpose and type. Additionally, 9.1% (N= 1) of them was not sure about it. With a mean score of 3.55/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other. Relatively, it was asked whether the stages of marking were reported and explained by means of the reliability estimates of the raters. At that point, it was reported that slightly higher than the half of the participants (N= 6; P= 54.6%) confirmed it. However, 45.5% (N= 5) of them were not sure whether the stages of marking were reported and explained profiting by raters' reliability estimates. With a mean score of 3.82/ 5.00 (SD= .65), the participants from the institution A held different opinions from each other. To note more, it was asked whether the criteria for marking were defined and readily available, such as rubrics, scales, schemes, answer keys, etc. In this context, almost all of the participants (N= 10; P= 90.9%) confirmed that above mentioned criteria and required rating scales were readily available. Not to mention, 9.1% (N= 1) of them were not sure whether those listed above were on hand. With a mean score of 4.00/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other.

With a view to marking and grading as ascertained by the ALTE, it was asked whether the data were collected from such a sample that could represent the candidates appropriately, regardless of any external factors like country of origin, gender, L1, age and so on and so forth. In this vein, it was reported that 63.7% (N= 7) of the participants confirmed the representativeness of the data for the candidates. However, 36.4% (N= 4) of them were not sure whether the data collected could identify the sample group of candidates adequately. With a mean score of 4.00/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other. Concomitantly, it was asked whether item-level data were collected from an appropriate sample of candidates in order to estimate item

difficulty, reliability measures, discrimination and standard errors of measurement. Accordingly, 72.7% (N= 8) of the participants confirmed that the collected item-level data could represent the sample of candidates adequately for the goodness of aforementioned estimates. Yet, 27.3% (N= 3) of them were not sure whether the data collected could make it possible to get item difficulty, discrimination, standard errors of measurement, and reliability estimates from an adequate sample of candidates. With a mean score of 3.82/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other.

With special interest to test analysis as ascertained by the ALTE, it was asked whether the characteristics of the test takers were defined clearly. Herein, slightly higher than the half of the participants (N= 6; P= 54.6%) confirmed that test takers' characteristics were purely defined. However, 27.3% (N= 3) of them were not sure whether the features of test takers were labelled obviously. Not to mention, 18.2% (N= 2) of them dissented to it, though. With a mean score of 3.64/ 5.00 (SD= .65), the participants from the institution A held different opinions from each other. At the very same, it was probed whether the tests were well-suited for test takers' overall abilities. In this context, 72.7% (N= 8) of the participants confirmed that the tests were pertinent for test takers concerning their overall abilities. Yet, 18.2% (N= 2) of them were not sure about it. Not to mention, 9.1% (N= 1) of them dissented to the fact that the tests verged upon the overall abilities of the test takers. With a mean score of 3.82/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other.

In addition to these, it was asked whether test takers were clued in the actual test formats. Herein, almost all of the participants (N= 10; P= 91.9%) confirmed the idea that test takers were well aware of the formats of the tests in use. Not to mention, 9.1% (N= 1) of them were not sure about it. With a mean score of 4.09/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other. Similarly, it was asked whether the tests in use were previously tried out on a sample population, which could be regarded as the representatives of the actual target test takers. Accordingly, there was an ambivalence between the ratios gained by the results as the participants who put forward that they were not sure whether previously mentioned was the case before administering the tests to the target population (N= 5; P= 45.5%) was the same with those who either agreed, or strongly

agreed it (N= 5; P= 45.5%). One more to note, 9.1% (N= 1) of the participants dissented to it, though. With a mean score of 3.55/ 5.00 (SD= .65), the participants from the institution A held different opinions from each other.

Correlatively, it was asked whether test results were reliable enough to get fair results. Herein, 45.5% (N= 5) of the participants were not sure whether test results could make them precipitate logical inferences. On the other hand, 36.4% (N= 4) of them confirmed that accurate decisions could be made thanks to the reliability of test results. Not to mention, 18.2% (N= 2) of them dissented to it. With a mean score of 3.36/ 5.00 (SD= .65), the participants from the institution A held different opinions from each other. Interestingly, when the degree of reliability gained by the numerical data was asked 63.6% (N= 7) of them confirmed that the reliability estimates were shown through statistics. Yet, 27.3% (N= 3) of them were not sure whether numerical data were supplied in order to demonstrate the reliability estimates of the tests in use. Besides, 9.1% (N= 1) of them directly dissented to it, though. With a mean score of 3.64/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other.

With respect to test analysis ascertained by the ALTE, it was also checked whether test formats were suitable and laced with clearly-defined contextual usages. Herein, 72.8% (N= 8) of the participants confirmed it; however, 18.2% (N= 2) of them were not sure whether that was case or not. In addition to these, 9.1% (N= 1) of the participants disagreed with it, though. With a mean score of 3.91/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other. Additionally, it was asked whether test formats could be transferred to the real testing situations. Herein, 72.8% (N= 8) of the participants confirmed it whereas 18.2% (N= 2) of them were not sure about it. Not to mention, 9.1% (N= 1) of them dissented to the fact that the test formats and features could be used in real testing situations, though. With a mean score of 3.82/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other. Moreover, it was asked whether the tests in use were applicable for the test item domain and the previously defined target population. In this vein, 72.8% (N= 8) of the participants confirmed it; yet, 27.3% (N= 3) of them were not sure about it. With a mean score of 3.82/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other. Last but not least, it was asked in relation to test analysis whether the proposed test

population and content showed similarity with the developmental sample. At that point, 63.6% (N= 7) of the participants confirmed the resemblance between the developmental sample and proposed test population. Nevertheless, 27.3% (N= 3) of them were not sure about it. To note more, 9.1% (N= 1) of them disagreed with it, though. With a mean score of 3.73/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other.

Within the scope of communication with stakeholders as ascertained by the ALTE, it was initially asked whether the test administration system could deliver the results to the candidates and exam centers swiftly, if required. Herein, 63.6% (N= 7) of the participants confirmed it whereas the rest (N= 4; P= 36.4%) was not sure whether the results were rendered promptly to the candidates and exam centers by the test administration system. With a mean score of 3.73/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other. Additionally, it was asked whether the stakeholders were given adequate information on the test purpose, content, use and reliability of the test results. In this context, 72.7% (N= 8) of the participants confirmed that the stakeholders were provided with information on above mentioned issues. However, 27.3% (N= 3) of them were not sure whether adequate information was given to the stakeholders related to topics aforementioned. With a mean score of 3.82/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other. Relatively, it was asked whether the stakeholders were given information on how to construe with the test results appropriately. Herein, higher than the half of the participants (N= 7; P= 63.6%) confirmed that they were informed about how to use the test results properly. Yet, 36.4% (N= 4) of them was not sure about it. With a mean score of 3.73/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other.

With respect to test production as ascertained by the ALTE, it was primarily asked whether it required great deal of training to conduct the tests. In this vein, 72.8% (N= 8) of the participants confirmed that training was needed to conduct the tests. However, 18.2% (N= 2) of them dissented to the fact that training was required before conducting the tests. Not to mention, 9.1% (N= 1) of them were not sure about it, though. With a mean score of 3.82/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other. Correlatively, it was asked whether it was easy to produce tests that were equivalent to the ones already in

use. Although the majority of the participants from the institution A asserted that great deal of training was required to conduct the tests, the 72.7% (N= 8) of them confirmed that it was easy to prepare equivalent test forms. Besides, 27.3% (N= 3) of them were not sure about it, though. Therefore, it could be stipulated that preparing the test forms did not take much time and necessitate training, but the stage of implementation did require so. With a mean score of 3.82/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other. Moreover, it was asked whether the tests in use were readily available. The majority of the participants from the institution A (N= 8; 72.7%) confirmed the presence of tests which were ready to be used for further implementations. Yet, 27.3% (N= 3) of them were not sure whether the tests were readily available. With a mean score of 4.00/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other.

In the same vein, it was also asked whether the tests were acceptable by the society and institutions. Herein, 63.6% (N= 7) of the participants confirmed it whereas 36.4% (N= 4) of them was not sure about the appropriateness of the tests by the society and institutions. With a mean score of 3.82/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other. At the very same, it was asked whether the tests were acceptable by the parents, teachers and administrators, as well. In this context, the majority of the participants (N= 9; P= 81.9%) confirmed that the tests in use were appropriate in the eyes of the aforementioned. Yet, 18.2% (N= 2) of them was not sure about it. With a mean score of 4.18/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other. Similarly, it was asked whether test takers were provided with different response items, such as sentence correction, gap filling, short answer, multiple choice, etc. Herein, slightly higher than the half of the participants (N= 6; P= 54.6%) confirmed that different response items were provided for test takers. However, 36.4% (N= 4) of them was not sure whether the test takers were supported with a variety of response items. Not to mention, 9.1% (N= 1) of them dissented to it, though. With a mean score of 3.82/ 5.00 (SD= .65), the participants from the institution A held different opinions from each other. One more to note, it was also asked whether the candidates were supplied with non-item based task types, such as speaking tasks, writing tasks and the like. In this vein, 81.9% (N= 9) of the

participants confirmed that non-item based task types were also provided for test takers. Yet, 18.2% (N= 2) of them was not sure whether such kind of task types were in use for the candidates. With a mean score of 4.18/ 5.00 (SD= .65), the participants from the institution A held similar opinion with each other.

The implementation of the ALTE code of practice by the institution B.

As previously mentioned, an overall estimation regarding the results gained from all of the private institutions were checked and reported together and separately. With this in mind, the overall results of the implementation of the ALTE Code of Practice are listed below regarding the case in private institution B.

With a view to test construction ascertained by the ALTE, it was initially checked whether the tests were grounded upon any theoretical construct. Concerning the results of the institution B, it was reported that almost all of the participants (N= 18; P= 94.8%) confirmed the basis of a pre-defined model, or a notional construct. On the other hand, 5.3% (N= 1) of the participants was not sure about it. With a mean score of 4.26/ 5.00 (SD= .56), the participants from the institution A held similar opinion with each other. Hence, it could be stipulated that participants from the institution B predominantly accepted the presence of a theoretical construct or a model applied within the tests. In the same vein, it was checked secondarily whether test purpose, its contextual use and population targeted at were defined appropriately. Concerning the results of the institution B, it was concluded that all of the participants (N= 19; P= 100%) confirmed that all aforementioned were defined properly. With a mean score of 4.47/ 5.00 (SD= .51), the participants from the institution B held the same opinion with each other. Relatively, 89.4% (N= 17) of the participants confirmed that the test purpose was clearly stated. On the other hand, 10.5% (N= 2) of them was not sure whether it was defined appropriately. With a mean score of 4.26/ 5.00 (SD= .65), the participants from the institution B held similar opinion with each other. By the same token, 84.2% (N= 16) of the participants from the institution B confirmed that the content of the tests was in correlation with the goal previously stated. However, 10.5% (N= 2) of them was not sure whether there was a correlation between the stated goal and test content. Besides, 5.3% (N= 1) of them dissented to the fact that there was a correlation between the stated goal and test content. With a mean score of 4.05/ 5.00 (SD= .78), the participants from the institution B held similar opinion with each

other. Similarly, it was also checked whether the tests encapsulated the full range of knowledge, skills and authentic language use, which could be applied in real world situations. Concerning the results of the institution B, it was yielded that 100% (N= 19) of the participants confirmed it, stating that the private institution they were working at (institution B) prepared tests that covered relevance among knowledge, skills and authenticity in terms of language use. With a mean score of 4.37/ 5.00 (SD= .50), the participants from the institution B held the same opinion with each other.

With a view to test construction ascertained by the ALTE, it was checked whether the tests in use were coupled with an external reference system, which was herein the Framework. Concerning the results of the institution B, it was yielded that 89.5% (N= 17) of the participants confirmed the relationship between the tests in use and CEFR. However, 10.5% (N= 2) of the participants were not sure whether an external criterion was followed and aligned while preparing the tests in use. With a mean score of 4.32/ 5.00 (SD= .67), the participants from the institution B predominantly confirmed the alignment of the tests to the Framework, holding similar opinion with each other. Uninterestingly, when they were asked about the evidences of the tests' linkage to the CEFR, 84.2% (N= 16) of the participants from the institution B were not sure whether there was an alignment chart available. On the other hand, 15.8% (N= 3) of them confirmed the presence of an alignment chart. With a mean score of 4.26/ 5.00 (SD= .73), the participants from the institution B held similar opinion with each other. In addition to these, it was also asked whether the tests in use were comparable with parallel examinations in recognition of content, consistency and parameters of grading. Herein, 84.2% (N= 16) of the participants confirmed the aforementioned comparability. However, 15.8% (N= 3) of them was not sure whether there was a comparison made amidst examinations parallel to each other within the scope of content, consistency and grading parameters. With a mean score of 4.26/ 5.00 (SD= .73), the participants from the institution B held similar opinion with each other. Besides, it was additionally checked whether there were criteria used for the selection and training of the test constructors, which were involved both in the processes of test construction and revision. In the light of this, 84.2% (N= 16) of the participants from the institution B confirmed that criteria for selection and training of test item writers were involved in

the processes aforementioned, and were also laced with expert judgment. However, 15.8% (N= 3) of them were not sure whether there were criteria pre-defined for test construction and revision of the tests. With a mean score of 4.26/ 5.00 (SD= .73), the participants from the institution B held similar opinion with each other. Moreover, the majority of the participants from the institution B (N= 16; P= 84.2%) supported that logical and empirical evidence were testified for the estimations of discriminant validity sub-scores. Yet, 10.5% (N= 2) of them was not sure whether these evidences were given. Additionally, 5.3% (N= 1) of them claimed that there was no any logical and empirical evidence given to support discriminant validity sub-scores. With a mean score of 4.05/ 5.00 (SD= .78), the participants from the institution B held similar opinion with each other.

With a view to administration and logistics ascertained by the ALTE, slightly higher than the half of the participants (N= 11; P= 57.9%) from the institution B confirmed that it did not cost a lot to administer the tests. On the other hand, 42.1% (N= 8) of the participants was not sure whether the tests in use were costly to procure and administer, composing the second largest proportion of all. With a mean score of 2.42/ 5.00 (SD= .51), the participants from the institution B held different opinions from each other. In the same vein, it was asked whether the test administration centers were selected concerning transparency with established set of procedures and regulations. Herein, 73.7% (N= 14) of the participants confirmed that the aforementioned centers selected accordingly. On the other hand, 21.1% (N= 4) of them were not sure whether these criteria were followed for the selection of test administration centers. Not to mention, 5.3% (N= 1) of them dissented to this fact, though. With a mean score of 3.79/ 5.00 (SD= .71), the participants from the institution B held similar opinion with each other. In addition to these, it was also checked whether test administration system had proper support systems. In this vein, the majority of the participants from the institution B (N= 12; P= 63.2%) confirmed that they had such support systems. On the other hand, 31.6% (N= 6) of them were not sure whether there were any support systems such as phone hotline or web services. One more to note, 5.3% (N= 1) of them stated that there were not any support systems as aforementioned. With a mean score of 3.79/ 5.00 (SD= .85), the participants from the institution B held similar opinion with each other. At that point, the participants were also asked whether the test administration system

pursued within the private institutions they were working at had support systems for the candidates with special needs. Herein, the majority of participants (N= 13; P= 68.5%) confirmed that there were support systems for the candidates with special needs. However, 26.3% (N= 5) were not sure whether the candidates with special needs were supported. Besides, 5.3% (N= 1) of the participants disagreed it. With a mean score of 3.84/ 5.00 (SD= .83), the participants from the institution B held similar opinion with each other.

With special interest to administration and logistics ascertained by the ALTE, it was also asked whether the results were protected appropriately by the security in order to keep the confidentiality of the test results and/or certificates. Herein, 68.5% (N= 13) of the participants confirmed that the security system in use could enable the confidentiality of the results. Yet, 26.3% (N= 5) of them were not sure whether the security support system could keep the confidentiality of the test results and/or certificates. One more to note, 5.3% (N= 1) of them were opposed to it, asserting that the results were not adequately protected. With a mean score of 3.84/ 5.00 (SD= .83), the participants from the institution B held similar opinion with each other. Correlatively, it was asked whether the tests were delivered to the scoring centers in good conditions thanks to security systems adopted. At that point, the majority of the participants from the institution B (N= 16; P= 84.2%) confirmed that the delivery was done properly and securely to the scoring centers. Not to mention, 15.8% (N= 3) of them were not sure whether security systems in use could make the test results delivered to the scoring centers appropriately. With a mean score of 3.95/ 5.00 (SD= .52), the participants from the institution B held similar opinion with each other.

Within the scope of marking and grading as ascertained by the ALTE, it was initially asked whether it was easy to score the tests, report the scores and interpret the results. Herein, nearly all of the participants from the institution B (N= 18; P= 94.7%) agreed upon it. Merely, 5.3% (N= 1) of them were not sure about it. Therefore, it could be stipulated that scoring the tests and interpreting the results were found easy together with reporting the scores. With a mean score of 4.21/ 5.00 (SD= .54), the participants from the institution B held similar opinion with each other. Additionally, it was also asked whether it was costly to score the tests. In this vein, there was a contradiction between the ones who stated that it was not costly to score

the tests and the ones who were not sure whether it costed too much to score the tests. To elaborate, slightly above than the half of the participants (N= 10; P= 52.6%) stated that it was not costly to score the tests. On the other hand, 47.4% (N= 9) of them were not sure whether it costed too much to score the tests. With a mean score of 2.47/ 5.00 (SD= .51), the participants from the institution B held different opinions from each other. Similarly, it was also asked whether marking was done accurately to yield reliable results in terms of test purpose and type. Herein, 79% (N= 15) of the participants confirmed that marking was adequately accurate and reliable. However, 15.8% (N= 3) of them were not sure about it. Additionally, 5.3% (N= 1) of the participants dissented to it, stating that marking was not sufficiently reliable for test purpose and type. With a mean score of 3.89/ 5.00 (SD= .74), the participants from the institution B held similar opinion with each other. Relatively, it was asked whether the stages of marking were reported and explained by means of the reliability estimates of the raters. At that point, it was reported that the majority of the participants (N= 15; P= 78.9%) confirmed it. However, 21.1% (N= 4) of them were not sure whether the stages of marking were reported and explained availing raters' reliability estimates. With a mean score of 4.05/ 5.00 (SD= .71), the participants from the institution B held similar opinion with each other. To note more, it was asked whether the criteria for marking were defined and readily available, such as rubrics, scales, schemes, answer keys, etc. In this context, almost all of the participants (N= 17; P= 89.5%) confirmed that above mentioned criteria and required rating scales were readily available. Not to mention, 10.5% (N= 2) of them were not sure whether those listed above were readily accessible. With a mean score of 4.00/ 5.00 (SD= .47), the participants from the institution B held similar opinion with each other.

With a view to marking and grading as ascertained by the ALTE, it was asked whether the data were collected from such a sample that could represent the candidates appropriately, regardless of any external factors like country of origin, gender, L1, age and so on and so forth. In this vein, it was reported that 73.7% (N= 14) of the participants confirmed the representativeness of the data for the candidates. However, 26.3% (N= 5) of them were not sure whether the data collected could typify the sample group of candidates adequately. With a mean score of 4.00/ 5.00 (SD= .75), the participants from the institution B held similar

opinion with each other. Concomitantly, it was asked whether item-level data were collected from an appropriate sample of candidates in order to estimate item difficulty, reliability measures, discrimination and standard errors of measurement. Accordingly, 73.7% (N= 14) of the participants confirmed that the collected item-level data could represent the sample of candidates adequately for the goodness of aforementioned estimates. Yet, 67.3% (N= 5) of them were not sure whether the data collected could make it possible to get item difficulty, discrimination, standard errors of measurement, and reliability estimates from an adequate sample of candidates. With a mean score of 3.89/ 5.00 (SD= .66), the participants from the institution B held similar opinion with each other.

With special interest to test analysis as ascertained by the ALTE, it was asked whether the characteristics of the test takers were defined clearly. Herein, nearly all of the participants (N= 17; P= 89.5%) confirmed that test takers' characteristics were apparently defined. However, 10.5% (N= 2) of them were not sure whether the features of test takers were labelled definitely. With a mean score of 4.16/ 5.00 (SD= .60), the participants from the institution B held similar opinion with each other. At the very same, it was probed whether the tests were convenient for test takers' overall abilities. In this context, 79% (N= 15) of the participants confirmed that the tests were applicable for test takers concerning their overall abilities. Yet, 21.1% (N= 4) of them were not sure about it. With a mean score of 4.11/ 5.00 (SD= .74), the participants from the institution B held similar opinion with each other. In addition to these, it was asked whether test takers were apprised of the substantive test formats. Herein, almost all of the participants (N= 17; P= 89.4%) confirmed the idea that test takers were well aware of the formats of the tests in use. Not to mention, 10.5% (N= 2) of them were not sure about it. With a mean score of 4.00/ 5.00 (SD= .47), the participants from the institution B held similar opinion with each other. Similarly, it was asked whether the tests in use were previously tried out on a sample population, which could be regarded as the representatives of the actual target test takers. Accordingly, almost all of the participants (N= 16; P= 89.4%) confirmed the idea that the tests were tried before administering them to the actual target population. Yet, 10.5% (N= 2) were not sure whether previously mentioned was the case before administering the tests to the target population. One more to note, 5.3% (N= 1) of the participants dissented to it, though. With a mean score of 4.16/ 5.00

(SD= .83), the participants from the institution B held similar opinion with each other. Correlatively, it was asked whether test results were reliable enough to get fair results. Herein, 84.3% (N= 16) of the participants confirmed that accurate decisions could be made thanks to the reliability of test results. On the other hand, 15.8% (N= 3) of them were not sure whether test results could make them precipitate logical inferences. With a mean score of 4.05/ 5.00 (SD= .62), the participants from the institution B held similar opinion with each other. Correlatively, when the degree of reliability gained by the numerical data was asked 73.7% (N= 14) of them confirmed that the reliability estimates were shown through statistics. Yet, 21.1% (N= 4) of them were not sure whether numerical data were supplied in order to demonstrate the reliability estimates of the tests in use. Besides, 5.3% (N= 1) of them directly dissented to it, though. With a mean score of 3.74/ 5.00 (SD= .65), the participants from the institution B held similar opinion with each other.

With respect to test analysis ascertained by the ALTE, it was also checked whether test formats were suitable and laced with clearly-defined contextual usages. Herein, 89.5% (N= 17) of the participants confirmed it; however, 5.3% (N= 1) of them were not sure whether that was case, or not. In addition to these, 5.3% (N= 1) of the participants disagreed with it, though. With a mean score of 4.16/ 5.00 (SD= .76), the participants from the institution B held similar opinion with each other. Additionally, it was asked whether test formats could be transferred to the real testing situations. Herein, 89.5% (N= 17) of the participants confirmed it whereas 10.5% (N= 2) of them were not sure about it. With a mean score of 4.11/ 5.00 (SD= .57), the participants from the institution B held similar opinion with each other. Moreover, it was asked whether the tests in use were applicable for the test item domain and the previously defined target population. In this vein, 89.5% (N= 17) of the participants confirmed it; yet, 10.5% (N= 2) of them were not sure about it. With a mean score of 4.05/ 5.00 (SD= .52), the participants from the institution B held similar opinion with each other. Last but not least, it was asked in relation to test analysis whether the proposed test population and content showed similarity with the developmental sample. At that point, 79% (N= 15) of the participants confirmed the resemblance between the developmental sample and proposed test population. Nevertheless, 21.1% (N= 4) of them were not sure about it. With a mean score of

4.11/ 5.00 (SD= .74), the participants from the institution B held similar opinion with each other.

Within the scope of communication with stakeholders as ascertained by the ALTE, it was initially asked whether the test administration system could deliver the results to the candidates and exam centers swiftly, if required. Herein, 68.5% (N= 13) of the participants confirmed it whereas the rest (N= 6; P= 31.6%) was not sure whether the results were rendered promptly to the candidates and exam centers by the test administration system. With a mean score of 3.89/ 5.00 (SD= .74), the participants from the institution B held similar opinion with each other. Additionally, it was asked whether the stakeholders were given adequate information on the test purpose, content, use and reliability of the test results. In this context, 79% (N= 15) of the participants confirmed that the stakeholders were provided with information on above mentioned issues. However, 21.1% (N= 4) of them were not sure whether adequate information was given to the stakeholders related to topics aforementioned. With a mean score of 3.95/ 5.00 (SD= .62), the participants from the institution B held similar opinion with each other. Relatively, it was asked whether the stakeholders were given information on how to construe with the test results appropriately. Herein, the majority of the participants (N= 15; P= 79%) confirmed that they were informed about how to use the test results properly. Yet, 21.1% (N= 4) of them were not sure about it. With a mean score of 4.00/ 5.00 (SD= .67), the participants from the institution B held similar opinion with each other.

With respect to test production as ascertained by the ALTE, it was primarily asked whether it required great deal of training to conduct the tests. In this vein, all of the participants (N= 19; P= 100%) confirmed that training was needed to conduct the tests. With a mean score of 4.32/ 5.00 (SD= .48), the participants from the institution B held the same opinion with each other. Correlatively, it was asked whether it was easy to produce tests that were equivalent to the ones already in use. Although the majority of the participants from the institution B asserted that great deal of training was required to conduct the tests, the 79% (N= 15) of them confirmed that it was easy to prepare equivalent test forms. Besides, 21.1% (N= 4) of them were not sure about it, though. Therefore, it could be stipulated that preparing the test forms did not take much time and necessitate training, albeit the stage of implementation did require so. With a mean score of 4.00/ 5.00 (SD= .67),

the participants from the institution B held similar opinion with each other. Moreover, it was asked whether the tests in use were readily available. The majority of the participants from the institution B (N= 10; P= 52.6%) was not sure about the presence of tests which were ready to be used for further implementations. Yet, 47.4% (N= 9) of them confirmed that the tests were readily available. With a mean score of 3.53/ 5.00 (SD= .61), the participants from the institution B held different opinions from each other.

In relation with the stage of test production as ascertained by the ALTE, it was also asked whether the tests were acceptable by the society and institutions. Herein, 84.2% (N= 16) of the participants confirmed it whereas 15.8% (N= 3) of them were not sure about the appropriateness of the tests by the society and institutions. With a mean score of 4.00/ 5.00 (SD= .58), the participants from the institution B held similar opinion with each other. At the very same, it was asked whether the tests were acceptable by the parents, teachers and administrators, as well. In this context, the majority of the participants (N= 14; P= 73.7%) confirmed that the tests in use were appropriate in the eyes of the aforementioned. Yet, 26.3% (N= 5) of them were not sure about it. With a mean score of 3.89/ 5.00 (SD= .66), the participants from the institution B held similar opinion with each other. In the same vein, it was asked whether test takers were provided with different response items, such as sentence correction, gap filling, short answer, multiple choice, etc. Herein, slightly higher than the half of the participants (N= 10; P= 52.6%) confirmed that different response items were provided for test takers. However, 42.1% (N= 8) of them were not sure whether the test takers were supported with a variety of response items. Not to mention, 5.3% (N= 1) of them dissented to it, though. With a mean score of 3.58/ 5.00 (SD= .77), the participants from the institution B held different opinions from each other. One more to note, it was also asked whether the candidates were supplied with non-item based task types, such as speaking tasks, writing tasks and the like. In this vein, 84.2% (N= 16) of the participants confirmed that non-item based task types were also provided for test takers. Yet, 15.8% (N= 3) of them were not sure whether such kind of task types were in use for the candidates. With a mean score of 3.89/ 5.00 (SD= .46), the participants from the institution B held similar opinion with each other.

The implementation of the ALTE code of practice by the institution C.

An overall estimation regarding the results gained from all of the private institutions were checked, and reported together and separately. With this in mind, the overall results of the implementation of the ALTE Code of Practice are listed below regarding the case in private institution C.

With a view to test construction ascertained by the ALTE, it was initially checked whether the tests were grounded upon any theoretical construct. Concerning the results of the institution C, it was reported that the majority of the participants (N= 6; P= 60%) confirmed the basis of a pre-defined model or a notional construct. On the other hand, 20% (N= 2) of the participants were not sure about it. Besides, 20% (N= 2) of them dissented to the fact that the tests in use were grounded upon a theory or a model. With a mean score of 3.30/ 5.00 (SD= 1.34), the participants from the institution C held similar opinion with each other. Hence, it could be stipulated that participants from the institution C populously accepted the presence of a theoretical construct or a model applied within the tests. Similarly, it was checked secondarily whether test purpose, its contextual use and population targeted at were defined appropriately. Concerning the results of the institution C, it was concluded that the majority of the participants (N= 6; P= 60%) confirmed that all aforementioned were defined properly. Yet, 40% (N= 4) of them disagreed it, though. With a mean score of 3.40/ 5.00 (SD= 1.26), the participants from the institution C held similar opinion with each other. Relatively, 70% (N= 7) of the participants confirmed that the test purpose was clearly stated. On the other hand, 20% (N= 2) of them were not sure whether it was defined appropriately. Besides, 10% (N= 1) of them disagreed it, though. With a mean score of 3.60/ 5.00 (SD= .57), the participants from the institution C held similar opinion with each other. By the same token, 70% (N= 7) of the participants from the institution C confirmed that the content of the tests was in correlation with the goal previously stated. However, 20% (N= 2) of them were not sure whether there was a correlation between the stated goal and test content. Besides, 10% (N= 1) of them dissented to the fact that there was a correlation between the stated goal and test content. With a mean score of 3.80/ 5.00 (SD= .54), the participants from the institution C held similar opinion with each other.

With a view to test construction ascertained by the ALTE, it was also checked whether the tests encapsulated the full range of knowledge, skills and authentic language use, which could be applied in real world situations. Concerning the results of the institution C, it was yielded that 60% (N= 6) of the participants confirmed it, stating that the private institution they were working at (institution C) prepared tests that covered relevance among knowledge, skills and authenticity in terms of language use. On the other hand, 20% (N= 2) of them were not sure whether that was the case, which was also backed up with the other 20% (N= 2) asserting that that was not the case. With a mean score of 3.20/ 5.00 (SD= .82), the participants from the institution C held similar opinion with each other. In the same vein, it was checked whether the tests in use were coupled with an external reference system, which was herein the Framework. Concerning the results of the institution C, it was yielded that 60% (N= 6) of the participants were not sure about such kind of a relationship between the tests in use and CEFR. However, 40% (N= 4) of the participants confirmed that an external criterion was followed and aligned while preparing the tests in use. With a mean score of 3.50/ 5.00 (SD= .71), the participants from the institution C predominantly confirmed the alignment of the tests to the Framework, holding similar opinion with each other.

In the same vein, when they were asked about the evidences of the tests' linkage to the CEFR, 60% (N= 6) of the participants from the institution C were not sure whether there was an alignment chart available. On the other hand, 40% (N= 4) of them confirmed the presence of an alignment chart. With a mean score of 3.40/ 5.00 (SD= .52), the participants from the institution C held different opinions from each other. In addition to these, it was also asked whether the tests in use were comparable with parallel examinations in recognition of content, consistency and parameters of grading. Herein, 30% (N= 3) of the participants confirmed the aforementioned comparability. However, 40% (N= 4) of them were not sure whether there was a comparison made amidst examinations parallel to each other within the scope of content, consistency and grading parameters. Besides, the rest (N=3; P= 30%) dissented to it, though. With a mean score of 3.00/ 5.00 (SD= .93), the participants from the institution C held different opinions from each other. Besides, it was additionally checked whether there were criteria used for the selection and training of the test constructors, which were involved both in the processes of test

construction and revision. In the light of this, half of the participants (N= 5; P= 50%) from the institution C confirmed that criteria for selection and training of test item writers were involved in the processes aforementioned, and were also laced with expert judgment. However, 30% (N= 3) of them were not sure whether there were criteria pre-defined for test construction and revision of the tests. Besides, 20% (N= 2) of them dissented to it, though. With a mean score of 3.30/ 5.00 (SD= .82), the participants from the institution C held different opinions from each other. Moreover, the majority of the participants from the institution C (N= 6; P= 60%) supported that logical and empirical evidences were testified for the estimations of discriminant validity sub-scores. Yet, 20% (N= 2) of them were not sure whether these evidences were given. Additionally, 20% (N= 2) of them claimed that there were not any logical and empirical evidences given to support discriminant validity sub-scores. With a mean score of 3.50/ 5.00 (SD= .61), the participants from the institution C held different opinions from each other.

With a view to administration and logistics ascertained by the ALTE, half of the participants (N= 5; P= 50%) from the institution C was not sure whether it costed a lot to administer the tests. On the other hand, 40% (N= 4) of the participants confirmed that the tests in use were costly to procure and administer, composing the second largest proportion of all. Additionally, 10% (N= 1) of them dissented to it, asserting that administering the tests was not that much costly. With a mean score of 3.40/ 5.00 (SD= .84), the participants from the institution C held different opinions from each other. Similarly, it was asked whether the test administration centers were selected concerning transparency with established set of procedures and regulations. Herein, 50% (N= 5) of the participants were not sure whether the aforementioned centers selected accordingly. On the other hand, 40% (N= 4) of them confirmed that these criteria were followed for the selection of test administration centers. Not to mention, 10% (N= 1) of them dissented to this fact, though. With a mean score of 3.40/ 5.00 (SD= .84), the participants from the institution C held different opinions from each other.

In addition to these, it was also checked whether test administration system had proper support systems. In this vein, the majority of the participants from the institution C (N= 6; P= 60%) confirmed that they had such support systems. On the other hand, 10% (N= 1) of them were not sure whether there were any support

systems such as phone hotline, or web services. One more to note, 30% (N= 3) of them stated that there were not any support systems as aforementioned. With a mean score of 3.20/ 5.00 (SD= 1.40), majority of the participants from the institution C were of the same opinion. At that point, the participants were also asked whether the test administration system pursued within the private institutions they were working at had support systems for the candidates with special needs. Herein, 40% (N= 4) of the participants from the institution C confirmed that there were support systems for the candidates with special needs. However, the same proportion of participants from the institution C (N= 4; P= 40%) stated that they were not sure whether the candidates with special needs were supported. Besides, 20% (N= 2) of the participants disagreed with it. With a mean score of 3.20/ 5.00 (SD= .79), the participants from the institution C held different opinions from each other.

To note more, it was also asked whether the results were protected appropriately by the security in order to keep the confidentiality of the test results and/or certificates. Herein, 40% (N= 4) of the participants claimed that the security system in use could not enable the confidentiality of the results. Yet, 30% (N= 3) of them was not sure whether the security support system could keep the confidentiality of the test results and/or certificates. One more to note, 30% (N= 3) of them confirmed that the results were adequately protected. With a mean score of 2.70/ 5.00 (SD= 1.16), the participants from the institution C held different opinions from each other. Correlatively, it was asked whether the tests were delivered to the scoring centers in good conditions thanks to security systems adopted. At that point, all of the participants from the institution C (N= 10; P= 100%) confirmed that the delivery was done properly and securely to the scoring centers. With a mean score of 4.10/ 5.00 (SD= .32), the participants from the institution C held the same opinion with each other.

Within the scope of marking and grading as ascertained by the ALTE, it was initially asked whether it was easy to score the tests, report the scores and interpret the results. Herein, all of the participants from the institution C (N= 10; P= 100%) agreed upon it. Therefore, it could be stipulated that scoring the tests and interpreting the results were found easy together with reporting the scores. With a mean score of 4.30/ 5.00 (SD= .48), the participants from the institution C were of the same opinion. Additionally, it was also asked whether it was costly to score the

tests. In this vein, slightly above than the half of the participants (N= 6; P= 60%) was not sure whether it was costly to score the tests. On the other hand, 30% (N= 3) of them confirmed that it costed too much to score the tests. Besides, 10% (N= 1) of them dissented to it, asserting scoring the tests did not cost a lot. With a mean score of 3.20/ 5.00 (SD= .63), the participants from the institution C held similar opinion with each other.

In the same vein, it was also asked whether marking was done accurately to yield reliable results in terms of test purpose and type. Herein, 50% (N= 5) of the participants confirmed that marking was adequately accurate and reliable. However, 20% (N= 2) of them were not sure about it. Additionally, 30% (N= 3) of the participants dissented to it, stating that marking was not sufficiently reliable for test purpose and type. With a mean score of 3.20/ 5.00 (SD= .92), the participants from the institution C held different opinions from each other. Relatively, it was asked whether the stages of marking were reported, and explained by means of the reliability estimates of the raters. At that point, it was reported that the majority of the participants (N= 6; P= 60%) was not sure about it. However, 20% (N= 2) of them confirmed that the stages of marking were reported and explained availing raters' reliability estimates. Besides, the other 20% (N= 2) disagreed with them, though. With a mean score of 2.80/ 5.00 (SD= 1.03), the participants from the institution C held similar opinion with each other.

To note more, it was asked whether the criteria for marking were defined and readily available, such as rubrics, scales, schemes, answer keys, etc. In this context, almost all of the participants (N= 7; P= 70%) confirmed that above mentioned criteria and required rating scales were readily available. Not to mention, 30% (N= 3) of them were not sure whether those listed above were readily accessible. With a mean score of 3.80/ 5.00 (SD= .63), the participants from the institution C held similar opinion with each other. Similarly, it was asked whether the data were collected from such a sample that could represent the candidates appropriately, regardless of any external factors like country of origin, gender, L1, age and so on and so forth. In this vein, it was reported that the estimates were dispersed in a two-way alternate. That was, the ratio of participants confirming the case aforementioned was proportionately the same with the ratio of participants who were not sure about it. To elaborate, 50% (N= 5) of the participants confirmed the

representativeness of the data for the candidates. However, 50% (N= 5) of them were not sure whether the data collected could typify the sample group of candidates adequately. With a mean score of 3.70/ 5.00 (SD= .80), the participants from the institution C held different opinions from each other.

Concomitantly, it was asked whether item-level data was collected from an appropriate sample of candidates in order to estimate item difficulty, reliability measures, discrimination and standard errors of measurement. Accordingly, 60% (N= 6) of the participants confirmed that the collected item-level data could represent the sample of candidates adequately for the goodness of aforementioned estimates. Yet, 40% (N= 4) of them were not sure whether the data collected could make it possible to get item difficulty, discrimination, standard errors of measurement, and reliability estimates from an adequate sample of candidates. With a mean score of 3.60/ 5.00 (SD= .52), the participants from the institution C held mostly similar opinion with each other.

With special interest to test analysis as ascertained by the ALTE, it was asked whether the characteristics of the test takers were defined clearly. Herein, the majority of the participants (N= 4; P= 40%) confirmed that test takers' characteristics were apparently defined. However, 30% (N= 3) of them were not sure whether the features of test takers were labelled definitely. Besides, 30% (N= 3) of them dissented to it, though. With a mean score of 3.10/ 5.00 (SD= 1.20), the participants from the institution C held different opinions from each other. At the very same, it was probed whether the tests were convenient for test takers' overall abilities. In this context, 40% (N= 4) of the participants confirmed that the tests were applicable for test takers concerning their overall abilities. Yet, the same proportion of participants (N= 4; P= 40%) of them disagreed with it, though. Additionally, 20% (N= 2) of them were not sure about any appropriateness, if that was the case over there. With a mean score of 3.20/ 5.00 (SD= 1.22), the participants from the institution C held different opinions from each other.

In addition to these, it was asked whether test takers were apprised of the substantive test formats. Herein, the majority of the participants (N= 7; P= 70%) confirmed the idea that test takers were well aware of the formats of the tests in use. Not to mention, 30% (N= 3) of them were not sure about it. With a mean score of 3.80/ 5.00 (SD= .65), the participants from the institution C held similar opinion with

each other. In the same vein, it was asked whether the tests in use were previously tried out on a sample population, which could be regarded as the representatives of the actual target test takers. Accordingly, 40% (N= 4) of the participants from the institution C confirmed the idea that the tests were tried before administering them to the actual target population. Yet, 30% (N= 3) of them were not sure whether previously mentioned was the case before administering the tests to the target population. One more to note, 30% (N= 3) of the participants dissented to it, though. With a mean score of 3.20/ 5.00 (SD= 1.03), the participants from the institution C held different opinions from each other.

Correlatively, it was asked whether test results were reliable enough to get fair results. Herein, 40% (N= 4) of the participants confirmed that accurate decisions could be made thanks to the reliability of test results. On the other hand, 40% (N= 4) of them disagreed with it, though. In addition to these, 20% (N= 2) of them were not sure whether test results could make them precipitate logical inferences. With a mean score of 2.80/ 5.00 (SD= 1.23), the participants from the institution C held different opinions from each other. Correlatively, when the degree of reliability gained by the numerical data was asked, 70% (N= 7) of them confirmed that the reliability estimates were shown through statistics. Yet, 20% (N= w) of them were not sure whether numerical data were supplied in order to demonstrate the reliability estimates of the tests in use. Besides, 10% (N= 1) of them directly dissented to it, though. With a mean score of 3.60/ 5.00 (SD= .70), the participants from the institution C held similar opinion with each other.

With respect to test analysis ascertained by the ALTE, it was also checked whether test formats were suitable and laced with clearly-defined contextual usages. Herein, half of the participants from the institution C (N= 5; P= 50%) of confirmed it; however, 20% (N= 2) of them were not sure whether that was the case, or not. In addition to these, 30% (N= 3) of the participants disagreed with it, though. With a mean score of 3.60/ 5.00 (SD= 1.35), the participants from the institution C were of different opinions from each other. Additionally, it was asked whether test formats could be transferred to the real testing situations. Herein, 40% (N= 4) of the participants confirmed it whereas the other 40% (N= 4) of them were not sure about it. Additionally, 20% (N= 2) of them dissented to it, though. With a mean score of

3.30/ 5.00 (SD= .95), the participants from the institution C held different opinions from each other.

Moreover, it was asked whether the tests in use were applicable for the test item domain and the previously defined target population. In this vein, 80% (N= 8) of the participants confirmed it; yet, 20% (N= 2) of them were not sure about it. With a mean score of 4.00/ 5.00 (SD= .67), the participants from the institution C held similar opinion with each other. Last but not least, it was asked in relation to test analysis whether the proposed test population and content showed similarity with the developmental sample. At that point, 50% (N= 5) of the participants confirmed the resemblance between the developmental sample and proposed test population. Nevertheless, 50% (N= 5) of them were not sure about it. With a mean score of 3.50/ 5.00 (SD= .53), the participants from the institution C held different opinions from each other.

Within the scope of communication with stakeholders as ascertained by the ALTE, it was initially asked whether the test administration system could deliver the results to the candidates and exam centers swiftly, if required. Herein, 50% (N= 5) of the participants confirmed it whereas 30% (N= 3) were not sure whether the results were rendered promptly to the candidates and exam centers by the test administration system. Besides, 20% (N= 2) of them disagreed with the rest, though. With a mean score of 3.30/ 5.00 (SD= .82), the participants from the institution C held different opinions from each other. Additionally, it was asked whether the stakeholders were given adequate information on the test purpose, content, use and reliability of the test results. In this context, 80% (N= 2) of the participants confirmed that the stakeholders were provided with information on above mentioned issues. However, 20% (N= 2) of them were not sure whether adequate information was given to the stakeholders related to topics aforementioned. With a mean score of 3.90/ 5.00 (SD= .57), the participants from the institution C held similar opinion with each other. Relatively, it was asked whether the stakeholders were given information on how to construe with the test results appropriately. Herein, the majority of the participants (N= 5; P= 50%) was not sure about what to do with the test results appropriately. On the other hand, 40% (N= 4) confirmed that they were informed about how to use the test results properly. Yet, 10% (N= 1) of them dissented to it, asserting that the stakeholders were not informed about the

interpretation of the results properly. With a mean score of 3.30/ 5.00 (SD= .67), the participants from the institution C held different opinions from each other.

With respect to test production as ascertained by the ALTE, it was primarily asked whether it required great deal of training to conduct the tests. In this vein, the majority of the participants (N= 7; P= 70%) confirmed that training was needed to conduct the tests. On the other hand, 20% (N= 2) of them were not whether training was needed to conduct the tests. Additionally, 10% (N= 1) of them claimed that not too much training was needed to conduct the tests. With a mean score of 4.00/ 5.00 (SD= 1.05), the participants from the institution C held similar opinion with each other. Correlatively, it was asked whether it was easy to produce tests that were equivalent to the ones already in use. Although the majority of the participants from the institution C asserted that great deal of training was required to conduct the tests, the 70% (N= 7) of them confirmed that it was easy to prepare equivalent test forms. Besides, 30% (N= 3) of them were not sure about it, though. Therefore, it could be stipulated that preparing the test forms did not take much time and necessitate training, albeit the stage of implementation did require so. With a mean score of 3.70/ 5.00 (SD= .48), the participants from the institution C held similar opinion with each other. Moreover, it was asked whether the tests in use were readily available. The majority of the participants from the institution C (N= 8; P= 80%) was not sure about the presence of tests which were ready to be used for further implementations. Yet, 10% (N= 1) of them confirmed that the tests were readily available. Besides, 10% (N= 1) of them disagreed with them, though. With a mean score of 3.00/ 5.00 (SD= .47), the participants from the institution C held similar opinion with each other.

In relation with the stage of test production as ascertained by the ALTE, it was also asked whether the tests were acceptable by the society and institutions. Herein, 70% (N= 7) of the participants confirmed it whereas 30% (N= 3) of them were not sure about the appropriateness of the tests by the society and institutions. With a mean score of 3.80/ 5.00 (SD= .63), the participants from the institution C held similar opinion with each other. At the very same, it was asked whether the tests were acceptable by the parents, teachers and administrators, as well. In this context, the majority of the participants (N= 7; P= 70%) confirmed that the tests in use were appropriate in the eyes of the aforementioned. Yet, 30% (N= 3) of them

were not sure about it. With a mean score of 4.00/ 5.00 (SD= .82), the participants from the institution C held similar opinion with each other. Similarly, it was asked whether test takers were provided with different response items, such as sentence correction, gap filling, short answer, multiple choice, etc. Herein, slightly higher than the half of the participants (N= 6; P= 60%) confirmed that different response items were provided for test takers. However, 30% (N= 3) of them were not sure whether the test takers were supported with a variety of response items. Not to mention, 10% (N= 1) of them dissented to it, though. With a mean score of 3.70/ 5.00 (SD= .95), the participants from the institution C held similar opinion with each other. One more to note, it was also asked whether the candidates were supplied with non-item based task types, such as speaking tasks, writing tasks and the like. In this vein, 90% (N= 9) of the participants confirmed that non-item based task types were also provided for test takers. Yet, 10% (N= 1) of them were not sure whether such kind of task types were in use for the candidates. With a mean score of 4.10/ 5.00 (SD= .57), the participants from the institution C held predominantly similar opinion with each other.

The overall picture of the implementation of the ALTE code of practice by selected private institutions. As previously mentioned, the ALTE Code of Practice were summed up in seven basic components within the questionnaire used for this study. These components were test construction, administration and logistics, marking and grading, test analysis, communication with stakeholders, test production, and item writing. Composed of 43 test items in total, these seven subsections were analyzed separately, and the estimations gained were reported singly.

Accordingly, in relation with the component of test construction, it could be stipulated that the tests in use were based on a theoretical construct, which was also laced with a well-defined purpose and context of use (N= 32; P= 80%). Furthermore, the tests were found directly associated with the purpose previously set (N= 29; P= 72.5%). By the same token, the tests in use were assumed to cover authentic use of the target language together with the required skills and knowledge (N= 33; P= 82.5%). Besides, it was yielded that test scores were in tune with an external criterion, which was, herein, accepted as the CEFR (N= 30; P= 75%). Thereby to note, when an alignment chart as a sign of the linkage between the test

scores and Framework was asked, a big proportion of participants was not well aware of it (N= 16; P= 40%), but the majority ascertained that they had it (N= 24; P= 60%). In parallel with this, it could be stipulated that there seemed to be a consistency across different examples of tests in use in recognition of content of use and boundaries of grading (N= 28; P= 70%). Herein, the results were assumed to be supported by some logical and empirical evidences, which were also corroborated with expert judgment in tow (N= 28; P= 70%).

With a view to administration and logistics, it could be stipulated that the tests were administered in the light of pre-defined set of regulations, laced with transparently established procedures and clear instructions (N= 25; P= 62.5%). However, the results yielded that administering the tests was found somehow costly as the majority of the participants from the selected private institutions stated so (N= 18; P= 45%). Apart from these, the tests were assumed to be delivered in good conditions and by secure means as the majority of the participants from the selected private institutions claimed so (N= 36; P= 90%). Nevertheless, when the support systems were asked, it could be stipulated by the results gained that slightly higher than the half of the participants confirmed the presence of such a system, which could be either a phone hotline, or web services provided (N= 22; P= 55%). Additionally, it was ascertained by the same ratio of participants that test takers with special needs were somehow supported (N= 22; P= 55%), but the rest was either not sure, or disagreed with it, though.

In relation with marking and grading, it could be stipulated that the majority of the participants was not well aware of how costly the scoring of the tests was (N= 16; P= 40%). However, the majority of them claimed that it was not difficult to score the tests and report the results afterwards (N= 39; P= 97.5%). Since, it was confirmed by most of the participants that marking schemes, rubrics and rating scales were easily on hand (N= 34; P= 85%). Moreover, the implementations related to marking were found sufficiently accurate by the majority of the participants (N= 7; P= 67.5%) as most of them certified that the marking procedure was documented and explained by means of reliability estimates of the raters at work (N= 23; P= 57.5%). When the reliability of the results was concerned, it could be stipulated that the results were purified of any external factors, such as L1, country of origin, age, gender, and the like (N= 26; P= 65%). One more to note, item-level data were

assumed to be gathered from an adequate sample of test takers in order to estimate the reliability, standard error of measurement and item difficulty accurately (N= 28; P= 70%).

With respect to test analysis, it was yielded by the results gained that the tests were appropriate to test takers' abilities to some extent (N= 27; P= 67.5%). Because it was predicated by the results gained that the features of the test takers were previously defined and labelled before test administration (N= 27; P= 67.5%). Besides, the format of the tests was found suitable with their contextual uses which were clearly defined (N= 30; P= 75%). Relatively, it was asserted by the majority of the participants that the test format could be easily applied in real testing settings (N= 29; P= 72.5%). That was why the test formats were assumed to be familiar for the proposed test takers by the majority of the participants (N= 34; P= 85%).

Moreover, it was questioned whether the tests in use were previously checked out on a sample which could be representative of the actual target population. Herein, the majority of the participants ascertained that the tests were piloted before they were administered to the target test takers (N= 25; P= 62.5%); however, one fourth of them was not sure about it (N= 10; P= 25%). The amount of the following made us infer that not all of the participants enrolled for this study were well-aware of a try-out before the actual implementation. Additionally, the tests in use were found relevant to the target population (N= 33; P= 82.5%). Correlatively, the test population proposed was found similar to the developmental sample (N= 27; P= 67.5%). Together with these, it could be indicated that the test results were found reliable enough to make accurate inferences by the majority of the participants (N= 24; P= 60%). However, one fourth of them was not sure about the reliability of the test results, though (N= 10; P= 25%). Within the scope of the reliability of the test results, it was also concluded that numerical data was provided to show the degree of reliability (N= 28; P= 70%).

With special concern upon communication with stakeholders, it was assumed that the stakeholders were informed on the tests, such as test purpose, content of use and test reliability (N= 31; P= 77.5%). Besides, it was also concluded that the stakeholders were informed about how to interpret and use the test results (N= 26; P= 65%). One more to note, it was noted that test administration system in use could communicate the results to the test takers in an accurate way (N= 25; P= 62.5%).

With regard to test analysis, it was reported that training was needed before administering the tests (N= 34; P= 85%). Interestingly, it was concluded that equivalent tests could be prepared easily (N= 30; P= 75%). However, when the participants were asked if the tests were readily available, most of them were not sure about it (N= 21; P= 52.5%), but there were some who confirmed that the tests were prepared beforehand and ready for use (N= 18; P= 45%). Apart from these, the tests were found not only societally and institutionally acceptable (N= 30; P= 75%), but also acceptable in the eyes of the administrators, parents and teachers, as well (N= 30; P= 75%).

Last but not least, in relation with item writing, it was concluded that the test takers were catered with different types of non-item based task types (N= 34; P= 85%). Besides, it was reported that the test takers were assumed to be provided with some types of response items, such as sentence correction, multiple choice, gap filling, short answer and the like (N= 22; P= 55%). However, it was also noted that the participants in no small measure were not sure about the fact that above mentioned types were in use within the tests produced (N= 15; P= 37.5%). Therefore, it could be stipulated that not all of the participants in this study, herein the English language teachers who were also working as test (-item) developers at those private institutions, were well-aware of the fact that test takers were provided with various types of response items, if that was the case.

Do the testing and assessment practices of non-formal English language schools in Turkey fit the guidelines assigned by ILTA? ILTA offers a number of basic tenets for its members by identifying the responsibilities of test designers, test item writers, institutions involved, stakeholders as the test result users, and test takers. Herein, ILTA buoys ethical standards by means of the 'Code of Ethics' (ILTA, 2000), and principles to enable good testing practice in all situations thanks to the 'Guidelines for Practice' (ILTA, 2007). In the light of these, the liabilities of the test developers at one side, and those of test takers on the other side are probed into below.

To elaborate, the items (N= 9) in the questionnaire regarding the ILTA Guidelines for Practice were numbered as 49, 50, 51, 52, 53, 54, 55, 56 and 57. The items were categorized into 2 sub-groups. These groups were named by ILTA itself as the 'responsibilities of the test designers and test writers, and the 'responsibilities

of the test takers'. As the guidelines set by ILTA reimbursed the ethical principles indeed, the Code of Ethics was not involved separately. The items in the questionnaire could be listed as given below:

Table 13

Questionnaire Items by the ILTA Guidelines for Practice

Section(s)	Sub-section(s)	Item(s)
The ILTA Guidelines for Practice	1. Responsibilities of the Test Designers and Test Writers (6 items)	Item No. 49: Test specifications and tasks are spelled out detail. Item No. 50: The tasks and test items are edited before (pre)testing. Item No. 51: The test materials are kept in a safe place. Item No. 52: Scoring procedures are carefully followed. Item No. 53: Items written by non-native speakers of the target language are checked by someone with a high-level of competence in the target language. Item No. 54: Test takers are treated with courtesy and respect during the testing process.
	2. Responsibilities of the Test Takers (3 items)	Item No. 55: Test takers read or listen to descriptive information and test instructions in advance of testing. Item No. 56: Test takers are well aware of the consequences of not taking the test. Item No. 57: Test takers can inform appropriate person(s), who are specified by the organization to be responsible for testing, if they believe that testing conditions have affected their results.
TOTAL	2 Sub-sections	9 Items

A sum of 9 items, which were above listed in detail, was taken to frequency analysis through descriptive statistics one by one. To add more, for each item, the participants' answers from 3 institutions were estimated and reported singly.

Accordingly, the first main consideration of ILTA, namely responsibilities of the test designers and test writers, was composed of 6 core items (item no. 49, 50, 51, 52, 53 and 54). This component was comprised of pre-defined test specifications and tasks, the safety of the test materials in use, the process of controlling the test items written, the on-going scoring procedures, and courtesy and respect aimed to be followed during the testing process. Each item was probed and described one by

one to give detailed information on the estimations gathered. Secondly, the responsibilities nestling test takers were checked with the help of 3 items (item no. 55, 56 and 57). It was initially asked whether test takers were provided with adequate information on the testing process. Besides, the consequences of not attending the testing process was questioned. One more to note, it was also examined whether test takers could have the privilege to inform someone from the authority on any condition that might affect the expected testing results. Before delving into details, table below given embodied the overall estimations regarding the exploitation of the ILTA Guidelines for Practice by selected private institutions. Means, standard deviations and standard errors of mean were given for each item elaborately.

Table 14

The Exploitation of the ILTA Guidelines for Practice by Selected Private Institutions

Section(s)	Item(s)	N	Mean	Std. Error of Mean	Std. Deviation
1. Responsibilities of the Test Designers and Test Writers	Item No. 49	40	4.05	.101	.638
	Item No. 50	40	4.10	.100	.632
	Item No. 51	40	4.13	.102	.648
	Item No. 52	40	4.18	.101	.636
	Item No. 53	40	3.90	.133	.841
	Item No. 54	40	3.95	.113	.714
2. Responsibilities of the Test Takers	Item No. 55	40	4.10	.086	.545
	Item No. 56	40	3.83	.129	.813
	Item No. 57	40	4.03	.091	.577
TOTAL	2 Sub-sections/ 9 Items	40			

In the light of these, it could be stipulated that the procedures concerning scoring of the tests were carefully proceeded as noted with the highest mean score of all within the scope of test designers' and test item writers' responsibilities (M= 4.18; SD= .64). It was followed by the assumption that the tests were kept safely with the second highest mean score of all (M= 4.13; SD= .65). Besides, it was also noted that the test items and task types went through the process of editing before administered to the target population (M= 4.10; SD= .63). Correlatively, the tasks and test specifications were marked to be unfolded in a clear way (M= 4.05; SD=

.64). Likewise, test takers were considered to be behaved in a respectful manner, and be acted in courtesy (M= 3.95; SD= .71). One more to note, it was certified by the participants of this study that the test items which were written by non-native speakers of the target language were presumably controlled by the authorities with a high-level of competence in the target language (M= 3.90; SD= .84).

Within the scope of the test takers' responsibilities, it was noted with the highest mean score of all that test takers had the opportunity to read or listen to instructions related to the testing procedure before the phase of implementation (M= 4.10; SD= .55). Besides, it was concluded that test takers somehow had the opportunity to inform the any authorized person during the phase of implementation about any problematic situation that could affect the reliability of the test results (M= 4.03; SD= .58). One more to note, it was also marked that test takers were cognizant of the results that might pop up if they happened to not take the test (M= 3.83; SD= .81).

Keeping these in mind, each sub-section was analyzed separately for each of the selected private institutions. The results were elaborated in detail, and the tables nestling all were given one by one. At first, an overall estimation regarding the results gained from all of the private institutions were checked and reported together. Following that, the results of each private institution were checked and reported separately by means of frequencies and percentages given within tables. With these in mind, the table below showed the overall results in a sub-section-based order before delving into the results of each private institution in detail. Each item was reported underneath singly, and the overall estimations were supported by their implications.

Table 15

The Implementation of the ILTA Guidelines for Practice by Selected Private Institutions

The implementation of the ILTA Guidelines for Practice		Strongly Disagree	Disagree	Not Sure	Agree	Strongly Agree	TOTAL
1. Item No. 49: Test specifications and tasks are spelled out detail.	f	0	0	7	24	9	40
	%	0.0	0.0	17.5	60.0	22.5	100
2. Item No. 50: The tasks and test items are	f	0	0	6	24	10	40

edited before (pre)testing.	%	0.0	0.0	15.0	60.0	25.0	100
3. Item No. 51: The test materials are kept in a safe place.	f	0	0	6	23	11	40
	%	0.0	0.0	15.0	57.5	27.5	100
4. Item No. 52: Scoring procedures are carefully followed.	f	0	0	5	23	12	40
	%	0.0	0.0	12.5	57.5	30.0	100
5. Item No. 53: Items written by non-native speakers of the target language are checked by someone with a high-level of competence in the target language.	f	0	1	13	15	11	40
	%	0.0	2.5	32.5	37.5	27.5	100
6. Item No. 54: Test takers are treated with courtesy and respect during the testing process.	f	0	0	11	20	9	40
	%	0.0	0.0	27.5	50.0	22.5	100
7. Item No. 55: Test takers read or listen to descriptive information and test instructions in advance of testing.	f	0	0	4	28	8	40
	%	0.0	0.0	10.0	70.0	20.0	100
8. Item No. 56: Test takers are well aware of the consequences of not taking the test.	f	0	2	11	19	8	40
	%	0.0	5.0	27.5	47.5	20.0	100
9. Item No. 57: Test takers can inform appropriate person(s), who are specified by the organization to be responsible for testing, if they believe that testing conditions have affected their results.	f	0	0	6	27	7	40
	%	0.0	0.0	15.0	67.5	17.5	100

The overall results above showed that 82.5% (N= 33) of the participants confirmed that test specifications were explained in great length. However, the rest (N= 7; P= 17.5%) was not sure whether that was the case over there. Additionally, 85% (N= 34) of the participants stated that the test items and tasks in use were brought into the phase of editing before administered to the target population. Yet, 15% (N= 6) of them were not sure whether editing was done before (pre)testing. A great majority of the participants (N= 34; P= 85%) also confirmed that the testing materials were kept safely and securely, as well. Moreover, in relation to the scoring procedure which was also labelled as one of the responsibilities of the test designers and/or test item writers, it was confirmed by the majority of the participants (N= 35; P= 87.5%) that scoring procedures were carried out with caution. Not to mention, 12.5% (N= 5) of them were not sure whether the procedures aforementioned were conducted carefully.

In the same vein, it was confirmed by the 65% (N= 26) of the participants that the test items which were prepared by the non-native speakers of the target

language were also reviewed by either native speakers of the target language, or someone proficient in the target language. Herein, 32.5% (N= 13) of the participants were not sure whether those test items underwent any preview within the private institution they were working at, though. Besides, it was questioned whether the candidates as the test takers were treated with respect during the implementation of the testing. Herein, it was noted that 72.5% (N= 29) of the participants confirmed it; however, 27.5% (N= 11) of them were not sure whether the test takers were behaved well thereat. Additionally, it was checked whether the test takers were catered with any instruction on how the testing process was followed. In this context, nearly all of the participants (N= 36; P= 90%) ascertained that such information was given to the candidates before testing was initiated. Not to mention, 10% (N= 4) of them were not sure about it, though.

Last but not least, the overall results above showed that 67.5% (N= 27) of the participants confirmed that all of the candidates were informed about the consequences of not sitting for the test. Herein, 27.5% (N= 11) of them were not sure whether the test takers were provided with information on the results of not taking the test. One more to note, 5% (N= 2) of them disagreed with them, asserting that the test takers were not informed about the case aforementioned, though. Once for all, the overall results above showed that the candidates were assumed to have the opportunity to call upon anyone responsible for testing on condition that any negative situation during testing might have an effect on the expected results (N= 34; P= 85%). Yet, 15% (N= 6) of them were not sure whether that was the case, or not.

Each of the private institutions was also checked separately to detect any implementational difference amidst. Accordingly, the results of each private institution were given below within tables embodied the estimations regarding the exploitation of the ILTA Guidelines for Practice by selected private institutions. For each item, frequencies and percentages were given within tables. The results were reported singly, and each item was elaborated in detail, embedding into sub-groups previously defined.

The implementation of the ILTA guidelines for practice by the institution

A. An overall estimation regarding the results gained from all of the private institutions were checked, and reported together and separately. With this in mind,

the overall results of the implementation of the ILTA Guidelines for Practice are listed below regarding the case in private institution A.

With a view to the 'responsibilities of the test designers and writers' ascertained by ILTA, it was checked initially whether test specifications and tasks in use were explained in detail. Concerning the results of the institution A, it was reported that the majority of the participants (N= 9; P= 81.8%) confirmed it. On the other hand, 18.2% (N= 2) of the participants were not sure whether the tasks and test specifications were spelled out comprehensively by the institution A. With a mean score of 4.11/ 5.00 (SD= .66), the participants from the institution A held similar opinion with each other. With a view to the 'responsibilities of the test designers and writers' ascertained by ILTA, it was secondarily checked whether the test items and tasks in use were edited before testing. Herein, the majority of the participants (N= 8; P= 72.8%) confirmed the editing phase whereas 27.3% (N= 3) of them were not sure whether editing was conducted before testing. With a mean score of 4.05/ 5.00 (SD= .62), the participants from the institution A held similar opinion with each other. In the same vein, it was also checked whether the test materials were kept carefully in order enable the confidentiality. In this context, almost all of the participants (N= 10; P= 91%) confirmed that the testing materials were preserved in safety. Yet, 9.1% (N= 1) of them were not sure about it, though. With a mean score of 4.05/ 5.00 (SD= .62), the participants from the institution A held similar opinion with each other.

With a view to the 'responsibilities of the test designers and writers' ascertained by ILTA, it was questioned whether scoring procedures were pursued delicately. Accordingly, all of the participants (N= 11; P= 100%) from the institution A either agreed, or strongly agreed with it. With no counter-view and a mean score of 4.00/ 5.00 (SD= .58), the participants from the institution A held the same opinion with each other. Similarly, it was probed whether the test items penned by the non-native speakers of the target language were reviewed by someone proficient in the target language, or the native speaker of the target language. Concordantly, it was noted that 63.7% (N= 7) of the participants confirmed it. However, 27.3% (N= 3) of them were not sure whether that was the case. Not to mention, 9.1% (N= 1) disagreed with it, though. With a mean score of 4.00/ 5.00 (SD= .82), the participants from the institution A held similar opinion with each other. Additionally, it was lastly

checked whether the test takers were behaved well in company with kindness and respect. In this vein, it was reported that the majority of the participants (N= 8; P= 72.8%) from the institution confirmed it, but the rest was not sure if it was the case (N= 3; P= 27.3%). With a mean score of 4.18/ 5.00 (SD= .87), the participants from the institution A held similar opinion with each other.

With a view to the 'responsibilities of the test takers' ascertained by ILTA, it was initially checked whether the test takers were given information on the testing process either visually or aurally in advance of testing. Herein, it was reported that 81.8% (N= 9) of the participants from the institution A confirmed it. On the other hand, 18.2 % (N= 2) of them were not sure whether the test takers were catered with explanatory instructions before taking the tests. With a mean score of 4.09/ 5.00 (SD= .87), the participants from the institution A held similar opinion with each other. Similarly, it was secondarily checked whether the test takers were given information on the probable sanctions of not taking the tests applied. At this point, 63.7% (N= 7) of the participants from the institution A confirmed that the candidates were well aware of the consequences aforementioned. Yet, 36.4% (N= 4) of them were not sure whether the test takers were provided with such an information beforehand. With a mean score of 3.91/ 5.00 (SD= .83), the participants from the institution A held similar opinion with each other.

Last but not least to report on the results of the institution A related to the 'responsibilities of the test takers' ascertained by ILTA, it was checked whether the test takers could inform someone authorized, if they happened to believe that the testing conditions were not appropriate enough, which could herewith yield negative effects on their results. Accordingly, it was noted that 81.8% (N= 9) of the participants confirmed it; however, 18.2% (N= 2) of them were not sure whether the test takers could have such an opportunity, or not. With a mean score of 4.00/ 5.00 (SD= .63), the participants from the institution A held similar opinion with each other.

The implementation of the ILTA guidelines for practice by the institution

B. An overall estimation regarding the results gained from all of the private institutions were checked, and reported together and separately. With this in mind, the overall results of the implementation of the ILTA Guidelines for Practice are given below regarding the case in private institution B.

With a view to the responsibilities of the test designers and writers ascertained by ILTA, it was checked initially whether test specifications and tasks in use were explained in detail. Concerning the results of the institution B, it was reported that the majority of the participants (N= 16; P= 84.2%) confirmed it. On the other hand, 15.8% (N= 3) of the participants were not sure whether the tasks and test specifications were spelled out comprehensively by the institution B. With a mean score of 4.11/ 5.00 (SD= .66), the participants from the institution B held similar opinion with each other. Similarly, it was secondarily checked whether the test items and tasks in use were edited before testing. Herein, the majority of the participants (N= 16; P= 84.2%) confirmed the editing phase whereas 15.8% (N= 3) of them were not sure whether editing was conducted before testing. With a mean score of 4.05/ 5.00 (SD= .62), the participants from the institution B held similar opinion with each other.

In the same vein, it was also checked whether the test materials were kept carefully in order enable the confidentiality. In this context, the majority of the participants (N= 16; P= 84.2%) confirmed that the testing materials were preserved in safety. Yet, 15.8% (N= 3) of them were not sure about it, though. With a mean score of 4.05/ 5.00 (SD= .62), the participants from the institution B held similar opinion with each other. Additionally, it was questioned whether scoring procedures were pursued delicately. Accordingly, the majority of the participants (N= 16; P= 84.2%) from the institution B either agreed, or strongly agreed with it. With some counter-views at the ratio of 15.8% (N= 3), and a mean score of 4.00/ 5.00 (SD= .58), the participants from the institution B held similar opinion with each other.

At the very same, it was probed whether the test items penned by the non-native speakers of the target language were reviewed by someone proficient in the target language, or the native speaker of the target language. Concordantly, it was noted that 68.4% (N= 13) of the participants confirmed it. However, 31.6% (N= 6) of them were not sure whether that was the case. With a mean score of 4.00/ 5.00 (SD= .82), the participants from the institution B held similar opinion with each other. One more to note, it was lastly checked whether the test takers were behaved well in company with kindness and respect. In this vein, it was reported that the majority of the participants (N= 13; P= 68.4%) from the institution confirmed it, but the rest

was not sure if it was the case (N= 6; P= 31.6%). With a mean score of 3.79/ 5.00 (SD= .63), the participants from the institution B held similar opinion with each other.

With a view to the responsibilities of the test takers ascertained by ILTA, it was initially checked whether the test takers were given information on the testing process either visually, or aurally in advance of testing. Herein, it was reported that 89.5% (N= 17) of the participants from the institution B confirmed it. On the other hand, 10.5% (N= 2) of them were not sure whether the test takers were catered with explanatory instructions before taking the tests. With a mean score of 4.11/ 5.00 (SD= .57), the participants from the institution B held similar opinion with each other. Similarly, it was secondarily checked whether the test takers were given information on the probable sanctions of not taking the tests applied. At this point, 78.9% (N= 15) of the participants from the institution B confirmed that the candidates were well aware of the consequences aforementioned. Yet, 21.1% (N= 4) of them were not sure whether the test takers were provided with such an information beforehand. With a mean score of 3.89/ 5.00 (SD= .57), the participants from the institution B held similar opinion with each other.

Last but not least to report on the results of the institution B related to the responsibilities of the test takers ascertained by ILTA, it was checked whether the test takers could inform someone authorized, if they happened to believe that the testing conditions were not appropriate enough, which could herewith yield negative effects on their results. Accordingly, it was noted that 84.2% (N= 16) of the participants confirmed it; however, 15.8% (N= 3) of them were not sure whether the test takers could have such an opportunity, or not. With a mean score of 4.00/ 5.00 (SD= .58), the participants from the institution B held similar opinion with each other.

The implementation of the ILTA guidelines for practice by the institution

C. An overall estimation regarding the results gained from all of the private institutions were checked, and reported together and separately. With this in mind, the table below showed the overall results of the implementation of the ILTA Guidelines for Practice in an item number-based order regarding the case in private institution C.

With a view to the responsibilities of the test designers and writers ascertained by ILTA, it was checked initially whether test specifications and tasks in

use were explained in detail. Concerning the results of the institution C, it was reported that the majority of the participants (N= 8; P= 80%) confirmed it. On the other hand, 20% (N= 2) of the participants were not sure whether the tasks and test specifications were spelled out comprehensively by the institution C. With a mean score of 3.90/ 5.00 (SD= .57), the participants from the institution C held similar opinion with each other.

Similarly, it was secondarily checked whether the test items and tasks in use were edited before testing. Herein, all of the participants (N= 10; P= 100%) confirmed that editing phase was conducted before testing with their replies of either 'agree' or 'strongly agree'. With no counter-view and a mean score of 4.10/ 5.00 (SD= .32), the participants from the institution C held the same opinion with each other. In the same vein, it was also checked whether the test materials were kept carefully in order enable the confidentiality. In this context, the majority of the participants (N= 8; P= 80%) confirmed that the testing materials were preserved in safety. Yet, 20% (N= 2) of them were not sure about it, though. With a mean score of 4.00/ 5.00 (SD= .67), the participants from the institution C held similar opinion with each other.

With a view to the responsibilities of the test designers and writers ascertained by ILTA, it was questioned whether scoring procedures were pursued delicately. Accordingly, the majority of the participants (N= 8; P= 80%) from the institution C either agreed, or strongly agreed with it. With some counter-views at the ratio of 20% (N= 2), and a mean score of 4.10/ 5.00 (SD= .74), the participants from the institution C held similar opinion with each other. Additionally, it was probed whether the test items penned by the non-native speakers of the target language were reviewed by someone proficient in the target language, or the native speaker of the target language. Concordantly, it was noted that 60% (N= 6) of the participants confirmed it. However, 40% (N= 4) of them were not sure whether that was the case. With a mean score of 3.80/ 5.00 (SD= .79), the participants from the institution C held different opinions from each other. One more to note, it was lastly checked whether the test takers were behaved well in company with kindness and respect. In this vein, it was reported that the majority of the participants (N= 8; P= 80%) from the institution confirmed it, but the rest was not sure if it was the case (N= 2; P=

20%). With a mean score of 4.00/ 5.00 (SD= .67), the participants from the institution C held similar opinion with each other.

With a view to the responsibilities of the test takers ascertained by ILTA, it was initially checked whether the test takers were given information on the testing process either visually or aurally in advance of testing. Herein, all of the participants (N= 10; P= 100%) from the institution C confirmed that the test takers were catered with explanatory instructions before taking the tests by their replies of either 'agree' or 'strongly agree'. With no counter-view and a mean score of 4.10/ 5.00 (SD= .32), the participants from the institution C held the same opinion with each other. Similarly, it was secondarily checked whether the test takers were given information on the probable sanctions of not taking the tests applied. At this point, half of the participants (N= 5; P= 50%) from the institution C confirmed that the candidates were well aware of the consequences aforementioned. Yet, 30% (N= 3) of them were not sure whether the test takers were provided with such an information beforehand. Not to mention, 20% (N= 2) of them dissented to it, though. With a mean score of 3.60/ 5.00 (SD= 1.17), the participants from the institution C held different opinions from each other.

Last but not least to report on the results of the institution C related to the responsibilities of the test takers ascertained by ILTA, it was checked whether the test takers could inform someone authorized, if they happened to believe that the testing conditions were not appropriate enough, which could herewith yield negative effects on their results. Accordingly, it was confirmed by almost all of the participants (N= 9; P= 90%) from the institution C that that was the case; however, 10% (N= 1) of them were not sure whether the test takers could have such an opportunity, or not. With a mean score of 4.10/ 5.00 (SD= .57), the participants from the institution C held similar opinion with each other.

The overall picture of the implementation of the ILTA guidelines for practice by selected private institutions. The ILTA Guidelines for Practice were summed up in two basic components within the questionnaire used for this study. These components were the 'responsibilities of the test designers and writers' and 'responsibilities of the test takers'. Composed of 9 test items in total, these two sub-sections were analyzed separately, and the estimations gained were reported singly.

Accordingly, the highest number of participants (N= 36) were estimated to either 'agree' or 'strongly agree' on the component of the responsibilities of the test takers, who confirmed at the ratio of 90% that the test takers were given information beforehand on what to do during the testing process. Hence, it could be inferred that the test takers were equipped with some prior knowledge on the process of testing before taking the tests. It was followed by the assumption that scoring procedures were followed carefully at the ratio of 87.5%. Therefore, it could be stipulated that the majority of the participants (N= 35) was convinced of the reliability of the scoring procedure conducted within the private institution they were working at. By the same token, it was postulated by the majority of the participants (N= 34) that the testing materials were ensured to be stored up in a safe place with the ratio of 85%. Likewise, it was reported by the majority of the participants (N= 34; P= 85%) that the candidates could acquaint someone responsible for the testing process with any distractive condition so as not to be effected by. Therefore, the test results could be more reliable and valid. The same ratio (P= 85%) also certified that the testing materials such as tasks, test items and the like underwent a process of editing before applied to the target population as an actual test. Herein, it could be asserted that the majority of the participants from the private institutions previously selected (N= 34) had a role in editing, as the participants of this study were not only English language teachers, but they were also the test (-item) developers at the private institutions they were working at.

Additionally, it was claimed by the majority of the participants (N= 33; P= 82.5%) that the tasks and specifications of the tests were elaborated so as not to be confusing for test takers by deviating from the actual test purpose(s). In the same vein, it was reported as one of the responsibilities of the test designers and test writers to make the test items typed by the non-native speakers of the target language controlled and revised by the native speakers of that language, or at least reviewed by someone with a high level of proficiency in that language. Herein, more than half of the participants (N= 26; P= 65%) confirmed that the review and/or revision procedure was followed. However, the ratio of participants who was not sure of it (N= 13; P= 32.5%), and that of disagreed ones (N= 1; P= 2.5%) could not be underestimated, as the number of participants was 40 in total. In the light of these, it could be stipulated that not all of the participants of this study, herein the

English language teachers who were also working as test (-item) developers at those private institutions, were well aware of such kind of implementations conducted within the institutions.

Besides, when it was questioned whether the test takers were treated with kindness and respect, it blossomed as a result that 72.5% (N= 29) of the participants confirmed it. However, there were 11 of them (P= 27.5%) who were not sure about this fact, though. Hence, it could be indicated that the one-fourth of the participants did not rest assured of the behaviors towards test takers. One more to note, although the number of participants (N= 27; P= 67.5%) who confirmed that the test takers were informed about the consequences of not taking the tests was higher than those who were not sure about it (N= 11; P= 27.5%) together with the number of participants who disagreed (N=2; P= 5%), the amount of the followings in tow could not be disregarded as there were 40 participants in total. Therefore, it could be speculated that not all of the participants of this study, the English language teachers who were also working as test (-item) developers at those private institutions, were well-aware of the fact that the candidates as the test takers were well informed upon the returns of not taking the tests.

What is the role of testing and assessment in Turkey's system of education in the light of the standards set by the AEA- Europe? Serving as a platform to create an environment of comparability across Europe within the scope of educational assessment, AEA- Europe has put forward the 'European Framework of Standards for Educational Assessment' (AEA- Europe, 2012) in order to provide information for test developers, score users and administrators as the educational authorities about the objectives of assessment, and about the claims of what the test results may probably mean. As it functions like a benchmarking criterion for already existing national system of specifications, it is more than a 'tool box' that requires strictly defined standards of performance in a technically detailed way. Therefore, the Framework reinforces practical implementations by means of a collection of appropriate methods in order to accomplish pre-defined requirements together with the best practices in reporting these methods. In doing these, it benefits from some 'guiding principles' and an 'instrument' which are also laced with a definition, purpose and a number of core elements in tow.

Within the scope of those guiding principles, the Framework suggests five tenets, which are deeply bounded to an educational purpose, and do fit for a European environment at the bottom. It also boosts fairness to keep the individual's rights through ethics. One more to note, it pinpoints practicality, validity and impact on stakeholders as a cornerstone for the review of the program, learning, test development and decision making. Correlatively, the instrument above mentioned fosters cooperation among standard requirements, methods and samples of evidence.

Taking these into consideration, the 'European Framework of Standards for Educational Assessment' (AEA- Europe, 2012) set by the AEA- Europe was used as a tool to answer this research question to define the general views and on-going practices of the selected private institutions on educational assessment in terms of two core components: (a) the guiding principles; and (b) the instrument to identify the nature of evidence, tasks and test types. To note beforehand, the items (N= 24) in the questionnaire regarding the European Framework of Standards for Educational Assessment set by AEA- Europe were numbered as 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81 and 82. The items were categorized into 2 groups. These groups were named by the AEA- Europe itself as the guiding principles and instrument for identifying the nature of evidence, tasks and test types. The items in the questionnaire could be listed as given below:

Table 16

Questionnaire Items of the European Framework of Standards for Educational Assessment by the AEA- Europe

Section(s)	Sub-section(s)	Item(s)
The European Framework of Standards for Educational Assessment by AEA- Europe	1. Guiding Principles (19 items)	Item No. 59: Overall evaluation of the total program, and assessment of educational systems are taken into consideration in testing procedures. Item No. 60: Innovative assessment techniques are taken into consideration while designing tests. Item No. 61: European perspective to the world-wide interest in assessment is adopted. Item No. 62: Establishing standards as a way of disseminating quality in assessment is the core element in testing and assessment practices. Item No. 63: The tests in use support different cultural and educational contexts.

1. Guiding Principles
(19 items)

Item No. 64: The test takers' place in the assessment process is well-defined.

Item No. 65: What is good for the individual in assessment aligns with the United Nations Convention on the Rights of the Child.

Item No. 66: Ethical considerations are given prominence in assessment procedures.

Item No. 67: The assessment belongs to the rights of the test takers; not to those who devise and administer the tests.

Item No. 68: The cornerstones of assessment (e.g. validity, practicality, impact on stakeholders) are carefully addressed.

Item No. 69: The results in the light of the essential quality aspects are meaningful and useful.

Item No. 70: Assessment translates the evidence that the results are defensible in different educational settings for further use.

Item No. 71: The purpose of the assessment supports the overall education of test takers.

Item No. 72: The assessment bases its rationale on the intended learning, which underlies a particular educational process.

Item No. 73: Assessment procedures provide information that confirms the aims of the Common European Framework of Reference for Languages.

Item No. 74: The kinds of assessment allow for feedback on the performance of the on-going educational system.

Item No. 75: Decision makers have the opportunity to evaluate programs and allocate resources by means of test results.

Item No. 76: The core elements of the Common European Framework of Reference for Languages are distinguished which follow the assessment development cycle.

Item No. 77: Possible evidences are presented to check whether the standard requirements are met by the test administered.

Item No. 78: The assessment applied in the institution/ organization covers standardized tests.

Item No. 79: The assessment applied in the institution/ organization covers school-based (summative) examinations.

Item No. 80: The assessment applied in the institution/ organization covers vocational (performance) assessment.

Item No. 81: The assessment applied in the institution/ organization covers learning outcomes of a curriculum (formative assessment).

Item No. 82: The assessment applied in the institution/ organization covers competency tests.

2. The Instrument
(5 items)

TOTAL

2 Sub-sections

24 Items

A sum of 24 items, which were above listed in detail, was taken to frequency analysis through descriptive statistics one by one. To add more, for each item, the participants' answers from 3 institutions were estimated and reported singly.

Accordingly, the first main consideration of the AEA- Europe, namely guiding principles, was composed of 19 core items (item no. 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76 and 77). Guiding principles were constituted by the overall evaluation of the total program by the testing procedures conducted, innovative assessment techniques in use, the European perspective adopted, the standards established to disseminate quality in assessment, the support given for variety of cultural and educational contexts, the definition of the test takers' place in the assessment process, some ethical considerations, the cornerstones of the assessment, the use of the assessment results for other educational settings, the rationale behind the assessment, the alignment of the test results to the CEFR, the dissemination of the results for further use, and possible evidences put forward as the standard requirements of the tests administered. Secondly, the instrument nestling the nature of evidence, tasks and test types were checked by means of 5 items (item no. 78, 79, 80, 81 and 82). Herein, it was asked what kind of tests were applied in practice by the previously selected private institutions, such as summative assessment, formative assessment, performance assessment, standardized tests and/or competency tests. Before delving into details, table below given embodied the overall estimations regarding the exploitation of the AEA- Europe's Framework of Standards by selected private institutions. Means, standard deviations and standard errors of mean were given for each item elaborately.

Table 17

The Exploitation of the AEA- Europe's Framework of Standards by Selected Private Institutions

Section(s)	Item(s)	N	Mean	Std. Error of Mean	Std. Deviation
1. Guiding Principles	Item No. 59	40	4.09	.163	.539
	Item No. 60	40	3.82	.182	.603
	Item No. 61	40	3.82	.296	.982
	Item No. 62	40	3.91	.163	.539
	Item No. 63	40	3.91	.211	.701
	Item No. 64	40	3.82	.226	.751
	Item No. 65	40	3.45	.207	.688
	Item No. 66	40	3.91	.251	.831

	Item No. 67	40	3.91	.251	.831
	Item No. 68	40	3.64	.310	1.03
	Item No. 69	40	4.09	.163	.539
	Item No. 70	40	4.00	.191	.632
	Item No. 71	40	3.73	.237	.786
1. Guiding Principles	Item No. 72	40	3.91	.163	.539
	Item No. 73	40	3.64	.310	1.03
	Item No. 74	40	4.27	.195	.647
	Item No. 75	40	3.91	.163	.539
	Item No. 76	40	3.91	.251	.831
	Item No. 77	40	3.64	.152	.505
	Item No. 78	40	3.45	.247	.820
2. The Instrument	Item No. 79	40	4.00	.191	.632
	Item No. 80	40	3.73	.237	.786
	Item No. 81	40	3.82	.122	.405
	Item No. 82	40	3.82	.122	.405
TOTAL	2 Sub-sections/ 24 Items	40			

In the light of these, the highest mean score related to the scope of guiding principles was the item asserting that assessment types were laced with feedback on the on-going educational system's overall performance ($M= 4.27$; $SD= .65$). Following that, the participants of this study stated that the test results could be appropriately used as one of the essentials of the quality as they were meaningful ($M= 4.09$; $SD= .54$). Likewise, it was stipulated by the results of this study that the testing procedures were adorned with the overall evaluation of the total program together with the assessment of the on-going educational system ($M= 4.09$; $SD= .54$). It was followed by the item claiming that the test results could be valid for various types of educational contexts for further use ($M= 4.00$; $SD= .63$). Alike, the tests in use were supposed to indorse the dissemination of the core principles of the actual testing and assessment practices ($M= 3.91$; $SD= .54$). At the very same, it was asserted by the participants of this study that decision makers had the opportunity to reckon with the programs through the exploitation of the test results ($M= 3.91$; $SD= .54$). Relatively, the assessment process was stipulated to have a basis on a rationale for the proposed learning of the predetermined educational process ($M= 3.91$; $SD= .54$). Moreover, the tests in use were supposed to cover various cultural and educational contexts ($M= 3.91$; $SD= .70$). In the same vein, the tests in use were assumed to be embellished with the elements of test development cycle of the CEFR ($M= 3.91$; $SD= .83$). Correlatively, the assessment procedures were alleged to consider some ethical concerns ($M= 3.91$; $SD= .83$), paying regard

not only to the rights of the test administrators, albeit to those of the test takers, as well (M= 3.91; SD= .83).

With regard to test design followed by the above-mentioned guiding principles, it was stipulated that innovative assessment techniques were considered in designing tests (M= 3.82; SD= .60). Correlatively, the assessment process was stipulated to cover the test takers' place within (M= 3.82; SD= .75). Besides, it was asserted that the tests in use were adorned with a European perspective to the assessment practices in a widespread interest (M= 3.82; SD= .98). Based on this, the purpose of the assessment was supposed to promote the overall education of the test takers (M= 3.73; SD= .79). To some extent, the anchors of the assessment process were assumed to be addressed delicately (M= 3.64; SD= 1.03). Furthermore, the assessment procedures were estimated to follow the aims set by the CEFR to some degree (M= 3.64; SD= 1.03). Herein, the tests in use were stipulated to cover the essentials of the assessment process by means of some possible evidences (M= 3.64; SD= .51). One more to note on the guiding principles of the AEA- Europe, it was stated that the rights of the test takers complied with the regulations of the 'United Nations Convention on the Rights of the Child' (UN, 1990) at the lowest ratio of mean score of all (M= 3.45; SD= .69).

Keeping these in mind, each sub-section was analyzed separately for each of the selected private institutions. The results were elaborated in detail, and the tables for each were given one by one. At first, an overall estimation regarding the results gained from all of the private institutions were checked and reported together. Following that, the results of each private institution were checked and reported separately by means of frequencies and percentages given within tables. With these in mind, the table below showed the overall results in a sub-section based order before delving into the results of each private institutions in detail. Each item was reported underneath singly, and the overall estimations were supported by their implications.

Table 18

The Implementation of the AEA- Europe's Framework of Standards by Selected Private Institutions

The Implementation of the AEA- Europe's Framework of Standards		Strongly Disagree	Disagree	Not Sure	Agree	Strongly Agree	TOTAL
1. Item No. 59: Overall evaluation of the total program, and assessment of educational systems are taken into consideration in testing procedures.	f	0	1	7	26	6	40
	%	0.0	2.5	17.5	65.0	15.0	100
2. Item No. 60: Innovative assessment techniques are taken into consideration while designing tests.	f	0	0	13	22	5	40
	%	0.0	0.0	32.5	55.0	12.5	100
3. Item No. 61: European perspective to the world-wide interest in assessment is adopted.	f	0	1	11	20	8	40
	%	0.0	2.5	27.5	50.0	20.0	100
4. Item No. 62: Establishing standards as a way of disseminating quality in assessment is the core element in testing and assessment practices.	f	0	0	9	25	6	40
	%	0.0	0.0	22.5	62.5	15.0	100
5. Item No. 63: The tests in use support different cultural and educational contexts.	f	0	1	9	22	8	40
	%	0.0	2.5	22.5	55.0	20.0	100
6. Item No. 64: The test takers' place in the assessment process is well-defined.	f	0	0	12	18	10	40
	%	0.0	0.0	30.0	45.0	25.0	100
7. Item No. 65: What is good for the individual in assessment aligns with the United Nations Convention on the Rights of the Child.	f	0	1	17	17	5	40
	%	0.0	2.5	42.5	42.5	12.5	100
8. Item No. 66: Ethical considerations are given prominence in assessment procedures.	f	0	0	13	22	5	40
	%	0.0	0.0	32.5	55.0	12.5	100
9. Item No. 67: The assessment belongs to the rights of the test takers; not to those who devise and administer the tests.	f	0	2	11	19	8	40
	%	0.0	5.0	27.5	47.5	20.0	100
10. Item No. 68: The cornerstones of assessment (e.g. validity, practicality, impact on stakeholders) are carefully addressed.	f	0	2	9	25	4	40
	%	0.0	5.0	22.5	62.5	10.0	100
11. Item No. 69: The results in the light of the essential quality aspects are meaningful and useful.	f	0	0	4	27	9	40
	%	0.0	0.0	10.0	67.5	22.5	100
12. Item No. 70: Assessment translates the evidence that the results are defensible in different educational settings for further use.	f	2	0	7	22	9	40
	%	5.0	0.0	17.5	55.0	22.5	100
13. Item No. 71: The purpose of the	f	0	1	4	31	4	40

assessment supports the overall education of test takers.	%	0.0	2.5	10.0	77.5	10.0	100
14. Item No. 72: The assessment bases its rationale on the intended learning, which underlies a particular educational process.	f	0	0	7	29	4	40
	%	0.0	0.0	17.5	72.5	10.0	100
15. Item No. 73: Assessment procedures provide information that confirms the aims of the Common European Framework of Reference for Languages.	f	1	0	7	29	3	40
	%	2.5	0.0	17.5	72.5	7.5	100
16. Item No. 74: The kinds of assessment allow for feedback on the performance of the on-going educational system.	f	0	0	7	22	11	40
	%	0.0	0.0	17.5	55.0	27.5	100
17. Item No. 75: Decision makers have the opportunity to evaluate programs and allocate resources by means of test results.	f	0	0	6	25	9	40
	%	0.0	0.0	15.0	62.5	22.5	100
18. Item No. 76: The core elements of the Common European Framework of Reference for Languages are distinguished which follow the assessment development cycle.	f	0	2	12	19	7	40
	%	0.0	5.0	30.0	47.5	17.5	100
19. Item No. 77: Possible evidences are presented to check whether the standard requirements are met by the test administered.	f	0	1	13	24	2	40
	%	0.0	2.5	32.5	60.0	5.0	100
20. Item No. 78: The assessment applied in the institution/ organization covers standardized tests.	f	1	1	9	21	8	40
	%	2.5	2.5	22.5	52.5	20.0	100
21. Item No. 79: The assessment applied in the institution/ organization covers school-based (summative) examinations.	f	1	0	11	22	6	40
	%	2.5	0.0	27.5	55.0	15.0	100
22. Item No. 80: The assessment applied in the institution/ organization covers vocational (performance) assessment.	f	1	0	12	21	6	40
	%	2.5	0.0	30.0	52.5	15.0	100
23. Item No. 81: The assessment applied in the institution/ organization covers learning outcomes of a curriculum (formative assessment).	f	0	2	7	29	2	40
	%	0.0	5.0	17.5	72.5	5.0	100
24. Item No. 82: The assessment applied in the institution/ organization covers competency tests.	f	0	1	6	30	3	40
	%	0.0	2.5	15.0	75.0	7.5	100

The overall results above showed that 80% (N= 32) of the participants confirmed that the testing procedures were framed by the overall evaluation of the program and on-going educational systems. On the other hand, 17.5% (N= 7) of the

participants was not sure whether the overall evaluation of the program and current education systems were taken into consideration while carrying out the testing procedures. Not to mention, 2.5% (N= 1) of the participants dissented to this fact although it was a drop in the ocean. Hence, it could be indicated that most of the participants accepted the presence of a validation of the testing procedures at the helm of the overall evaluation of the total program together with the current educational systems in use. Besides, it was asserted as one of the core elements of the testing and assessment practices that some standards were to be set in order to disseminate quality in them. In relation with this, 77.5% (N= 31) of the participants confirmed it. Yet, the rest as the 22.5% (N= 9) of them were not sure whether that was the case in practice. In doing this, the tests in use were stipulated to cover wide range of cultural and educational contexts by the three-quarter of the overall participants (N= 30; P= 75%). However, 22.5% (N= 9) of them were not sure whether variation in contextual usage of the target language was addressed by the tests in use. Not to mention, 2.5% (N= 1) of them dissented to it, though.

Furthermore, it was asserted by the majority of the participants (N= 28; P= 70%) that European perspective was pursued while designing tests, and developing test items. Yet, 27.5% (N= 11) of them were not sure about it. Additionally, 2.5% (N= 1) of them disagreed with them, though. Correlatively, it was confirmed by the majority of the participants (N= 27; P= 67.5%) that innovative assessment techniques were used by the private institutions previously selected. However, slightly above than the three-quarter of the participants (N= 13; P= 32.5%) was not sure whether innovative techniques were adopted in testing and assessment practices. At the very same, 70% (N= 28) of the participants supported that the assessment process did not run independently, instead took the place of the test takers within the testing procedure into consideration. On the other hand, 30% (N= 12) of them were not sure whether test takers' place was properly defined in the assessment process.

That was why it was confirmed by the majority of the participants (N= 27; P= 67.5%) that the test takers' rights were concerned in the assessment process. Yet, 27.5% (N= 11) of them were not sure whether they were the rights of the ones who did devise and administer the tests which were protected, albeit not the ones who took the tests as the candidates. Not to mention, 5% (N= 2) of them disagreed with

it, though. Herein, it was reported that slightly higher than the half of the participants (N= 22; P= 55%) confirmed the alignment of the individual's place in the assessment procedure with the United Nations Convention on the Rights of the Child. However, nearly half of them (N= 17; P= 42.5%) was not sure about it. Besides, 2.5% (N= 1) of them disagreed with them, though. In relation to this, 67.5% (N= 27) of the participants confirmed that ethical issues were taken into consideration within assessment procedures. On the other hand, 32.5% (N= 13) of them were not sure whether that was the case.

Moreover, it was confirmed by the 72.5% (N= 29) of the participants that the essentials of assessment procedure were marked properly. Yet, 22.5% (N= 9) of them were not sure whether that was the case. Not to mention, 5% (N= 2) of them dissented to it, though. Additionally, a great majority of the participants (N= 31; P= 90%) asserted that the test results were functional and purposeful although 10% (N= 4) of them were not sure about it. Correlatively, 77.5% (N= 31) of the participants confirmed that the test results could be utilized in other educational settings for further use. But 17.5% (N= 7) of them were not sure about it in addition to other 5% (N= 2) stating that the test results could not be used afterwards for any educational purpose as they strongly disagreed with it. Additionally, 87.5% (N= 35) of the participants confirmed that the overall education of the test takers was supported by the tests in use. Yet, 10% (N= 4) of them were not sure about it. Not to mention, 2.5% (N= 1) of them dissented to it, though. Concomitantly, 82.5% (N= 33) of the participants confirmed that the test takers were provided with tests which could assign its rationale on the intended learning in tow. Yet, 17.5% (N= 7) of the participants were not sure about it. To note more, 80% (N= 32) of the participants confirmed that the tests in use were correlated with the rudiments of the CEFR. However, 17.5% (N= 7) of them were not sure whether the tests in use provided information on the adoption of the objectives set by the CEFR. Besides, 2.5% (N= 1) of them strongly disagreed with the others, though.

Furthermore, 65% (N= 26) of the participants certified that the basic tenets of the CEFR were pursued by the test development cycle. Yet, 30% (N= 12) of them were not sure about it. In addition to them, 5% (N= 2) of the participants disagreed with it, though. Besides, the majority of the participants (N= 33; P= 82.5%) confirmed that there were different types of assessment in use which enabled getting feedback

on the on-going educational system, as well. Yet, 17.5% (N= 7) of them were not sure about it. Relatively, 85% (N= 34) of the participants confirmed that different assessment types in use together with the test results gained allowed for the evaluation of the program by decision-makers. However, 15% (N= 6) of them were not sure whether the decision-makers could have the opportunity to evaluate the current program by means of the test results gained. Herein, it was also confirmed by the 65% (N= 26) of the participants that potential evidences were given after the implementation of each test in order to check whether the standard requirements previously defined were met, or not. Yet, 32.5% (N= 13) of them were not sure about it. Not to mention, 2.5% (N= 1) of them disagreed with it, though.

In relation with the component of instrument, the overall results above showed that 72.5% (N= 29) of the participants confirmed the implementation of standardized tests within their institutions. Yet, 22.5% (N= 9) of them were not sure whether standardized tests were in use. Besides, 5% (N= 2) of them dissented to the use of standardized tests within their institutions, though. Correlatively, 70% (N= 28) of the participants confirmed that school-based examinations were applied within the previously selected private institutions. On the other hand, 27.5% (N= 11) of them were not sure whether school-based examinations were in use as a part of summative assessment. Besides, 2.5% (N= 1) of them strongly disagreed with them, though.

To note more, 67.5% (N= 27) of the participants stated that performance assessment was conducted as an implementation within the previously selected private institutions. Yet, 30% (N= 12) of them were not sure whether that was the case. Besides, 2.5% (N= 1) of them strongly disagreed with them, though. Within the scope of assessment practices conducted by the selected private institutions, 77.5% (N= 31) of the participants confirmed that formative assessment was conducted within the previously selected private institutions. Yet, 17.5% (N= 7) of them were not sure whether learning outcomes of a curriculum were covered by means of formative assessment. Besides, 5% (N= 2) of them disagreed with them, though. To mark as a last item for the 'instrument', it was confirmed by the majority of the participants (N= 33; P= 82.5%) that competency tests were in use within the previously selected private institutions. Yet, 15% (N= 6) of them were not sure

whether that was the case. Not to mention, 2.5% (N= 1) of them disagreed with them, though.

Each of the private institutions was also checked separately to detect any implementational difference amidst. Accordingly, the results of each private institution were given below by the frequencies and percentages. The results were reported singly, and each item was elaborated in detail, embedding into subgroups previously defined.

The implementation of the AEA- Europe's framework of standards by the institution A. An overall estimation regarding the results gained from all of the private institutions were checked, and reported together and separately. With this in mind, the overall results of the implementation of the AEA- Europe's Framework of Standards are given below regarding the case in private institution A.

With a view to the guiding principles ascertained by the AEA- Europe, it was initially checked whether the testing procedures were framed by the overall evaluation of the program and on-going educational systems. Concerning the results of the institution A, it was reported that almost all of the participants (N= 10; P= 90.9%) confirmed taking the overall evaluation of the program and current education systems into consideration while carrying out the testing procedures. On the other hand, 9.1% (N= 1) of the participants were not sure about it. With a mean score of 4.09/ 5.00 (SD= .54), the participants from the institution A held similar opinion with each other. Hence, it could be stipulated that participants from the institution A predominantly accepted the presence of a validation of the testing procedures at the helm of the overall evaluation of the total program together with the current educational systems in use.

With a view to the guiding principles ascertained by the AEA- Europe, it was secondarily checked whether any standards were set before administering the tests in order to disseminate quality in related testing and assessment practices. In relation with this, 81.8% (N= 9) of the participants confirmed it. Yet, the rest as the 18.2% (N= 2) of them were not sure whether that was the case in practice. With a mean score of 3.91/ 5.00 (SD= .54), the participants from the institution A held similar opinion with each other. Therefore, it could be stipulated that some core elements were defined as the standards of testing and assessment practices before

the tests were administered to the target population. In the same vein, it was also probed whether the tests in use covered a wide range of cultural and educational contexts within. Herein, it was reported by the majority of the participants (N= 8; P= 72.7%) that the tests in use were constituted by various cultural and educational contexts within. However, 27.3% (N= 3) of them were not sure whether variation in contextual usage of the target language was addressed by the tests in use. With a mean score of 3.91/ 5.00 (SD= .70), the participants from the institution A held similar opinion with each other. Similarly, it was checked whether a European perspective was pursued while designing tests, and developing test items. Herein, slightly higher than the half of the participants (N= 6; P= 54.5%) was not sure whether that was the case. On the other hand, 45.5% (N= 5) of them confirmed adopting a European perspective in designing tests. With a mean score of 3.82/ 5.00 (SD= .98), the participants from the institution A held different opinions from each other. Correlatively, it was scrutinized whether innovative assessment techniques were used by the private institutions previously selected. In this context, 72.7% (N= 8) of the participants confirmed it. However, 27.3% (N= 3) of the participants were not sure whether innovative techniques were adopted in testing and assessment practices. With a mean score of 3.82/ 5.00 (SD= .60), the participants from the institution A held similar opinion with each other.

With a view to the guiding principles ascertained by the AEA- Europe, it was also checked whether the assessment process did run taking the place of the test takers within the testing procedure into consideration. Herein, 63.7% (N= 7) of the participants from the institution A confirmed that the assessment process did not run independently, instead took the place of the test takers within the testing procedure into consideration. Yet, 36.4% (N= 4) of them were not sure about it. With a mean score of 3.82/ 5.00 (SD= .75), the participants from the institution A were of similar opinion with each other. Concomitantly, it was questioned whether the test takers' rights were concerned in the assessment process. Herein, 63.7% (N= 7) of the participants confirmed it. But 36.4% (N= 4) of them were not sure whether they were the rights of the ones who did devise and administer the tests which were protected, albeit not the ones who took the tests as the candidates. With a mean score of 3.91/ 5.00 (SD= .83), the participants from the institution A were of similar opinion with each other.

Relatively, it was checked whether the alignment of the individual's place in the assessment procedure with the United Nations Convention on the Rights of the Child was available. In this vein, the majority of the participants (N= 7; P= 63.6%) from the institution A was not sure whether that was the case. On the other hand, 36.4% (N= 4) of them confirmed it. With a mean score of 3.45/ 5.00 (SD= .69), the participants from the institution A held similar opinion with each other. Therefore, it could be stipulated that the majority of the participants from the institution A was not well aware of the presence of such an alignment between the test takers' rights and the United Nations Convention on the Rights of the Child. To note more, it was scrutinized whether ethical issues were taken into consideration within assessment procedures. Herein, 63.7% (N= 7) of the participants from the institution A confirmed it. Yet, 36.4% (N= 4) of them were not sure whether that was the case. With a mean score of 3.91/ 5.00 (SD= .83), the participants from the institution A held similar opinion with each other.

With a view to the guiding principles ascertained by the AEA- Europe, it was checked whether the essentials of assessment procedure were marked properly. In this case, 63.7% (N= 7) of the participants from the institution A confirmed that the keystones of assessment were properly addressed. On the other hand, 18.2% (N= 2) of them were not sure about it. Besides, 18.2% (N= 2) of them dissented to it, though. With a mean score of 3.64/ 5.00 (SD= 1.03), the participants from the institution A held similar opinion with each other. Additionally, it was questioned whether the test results were functional and purposeful. Herein, almost all of the participants (N= 9; P= 90.9%) from the institution A confirmed it. Yet, 9.1% (N= 1) of them were not sure about it. With a mean score of 4.09/ 5.00 (SD= .54), the participants from the institution A held similar opinion with each other.

Correlatively, it was checked whether the test results could be utilized in other educational settings for further use. Herein, 81.8% (N= 9) of the participants from the institution A confirmed it in contrast with the 18.2% (N= 2) of them, stating that they were not sure whether the test results could be used afterwards for any other educational purpose. With a mean score of 4.00/ 5.00 (SD= .63), the participants from the institution A held similar opinion with each other. In the same vein, it was also scrutinized whether the overall education of the test takers was supported by the tests in use. In this context, 72.7% (N= 8) of the participants confirmed it. Yet,

18.2% (N= 2) of them were not sure about it. In addition to these, 9.1% (N= 1) of them dissented to it, though. With a mean score of 3.73/ 5.00 (SD= .79), the participants from the institution A held similar opinion with each other. Concomitantly, it was delved whether that the test takers were provided with tests which could assign its rationale on the intended learning in tow. Herein, the majority of the participants from the institution A (N= 9; P= 81.8%) confirmed it whereas 18.2% (N= 2) of them were not sure about it. With a mean score of 3.91/ 5.00 (SD= .54), the participants from the institution A held similar opinion with each other.

With reference to the guiding principles ascertained by the AEA- Europe, it was checked whether the tests in use were correlated with the rudiments of the CEFR. Herein, 72.7% (N= 8) of the participants from the institution A confirmed it. However, 18.2% (N= 2) of them were not sure whether the tests in use provided information on the adoption of the objectives set by the CEFR. Not to mention, 9.1% (N= 1) of them strongly disagreed with the other, though. With a mean score of 3.64/ 5.00 (SD= 1.03), the participants from the institution A held similar opinion with each other. Additionally, it was questioned whether the basic tenets of the CEFR were pursued by the test development cycle. Accordingly, more than half of the participants (N= 7; P= 63.7%) confirmed it. On the other hand, 36.4% (N= 4) of them were not sure whether that was the case. With a mean score of 3.91/ 5.00 (SD= .83), the participants from the institution A held similar opinion with each other. Similarly, it was also scrutinized whether there were different types of assessment in use which enabled getting feedback on the on-going educational system. Accordingly, almost all of the participants (N= 10; P= 90.9%) from the institution A confirmed it in spite of the 9.1% (N= 1) of them who were not sure about it. With a mean score of 4.27/ 5.00 (SD= .65), the participants from the institution A held very similar opinion with each other.

Correlatively, it was checked whether different assessment types in use together with the test results gained allowed for the evaluation of the program by decision-makers. Herein, 81.8% (N= 9) of the participants from the institution A confirmed that the decision-makers could have the opportunity to evaluate the current program by means of the test results gained. Yet, 18.2% (N= 2) of them were not sure whether that was the case. With a mean score of 3.91/ 5.00 (SD= .54), the participants from the institution A held similar opinion with each other.

Additionally, it was probed whether potential evidences were given after the implementation of each test in order to check whether the standard requirements previously defined were met, or not. Herein, 63.6% (N= 7) of the participants confirmed it whereas 36.4% (N= 4) of them were not sure about it. With a mean score of 3.64/ 5.00 (SD= .50), the participants from the institution A held similar opinion with each other.

With special reference to the instrument ascertained by the AEA- Europe, the assessment types in use as an instrument for testing and assessment practices were questioned. Accordingly, there was a contradiction between the ones confirming the exploitation of standardized tests (N= 5; P= 45.5%) and the ones who were not sure whether standardized tests were in use by the institution A (N= 5; P= 45.5%). Besides, 9.1% (N= 1) of them disagreed that standardized tests were conducted within the institution A. With a mean score of 3.45/ 5.00 (SD= .82), the participants from the institution A held different opinions from each other. At the very same, it was certified by the majority of the participants (N= 9; P= 81.8%) that summative assessment such as school-based examinations were applied within the institution A. Yet, 18.2% (N= 2) of them were not sure whether such an assessment type was being conducted. With a mean score of 4.00/ 5.00 (SD= .63), the participants from the institution A held similar opinion with each other.

Apart from these, it was asserted by slightly higher than the half of the participants (N= 6; P= 54.6%) that performance assessment was conducted as an implementation within the institution A. However, nearly other half of the participants (N= 5; P= 45.5%) was not sure whether that was the case. With a mean score of 3.73/ 5.00 (SD= .79), the participants from the institution A held different opinions from each other. To add more, the majority of the participants (N= 9; P= 81.8%) from the institution A confirmed that formative assessment was conducted within. Yet, 18.2% (N= 2) of them were not sure whether learning outcomes of a curriculum were covered by means of formative assessment. With a mean score of 3.82/ 5.00 (SD= .40), the participants from the institution A held similar opinions with each other. Likewise, the majority of the participants (N= 9; P= 81.8%) from the institution A confirmed that competency tests were in use within the institution A. Yet, 18.2% (N= 2) of them were not sure whether that was the case. With a mean score of 3.82/

5.00 (SD= .40), the participants from the institution A held similar opinion with each other.

The implementation of the AEA- Europe's framework of standards by the institution B. An overall estimation regarding the results gained from all of the private institutions were checked, and reported together and separately. With this in mind, the overall results of the implementation of the Framework of Standards by the AEA-Europe are listed below regarding the case in private institution B.

With a view to the guiding principles ascertained by the AEA- Europe, it was initially checked whether the testing procedures were framed by the overall evaluation of the program and on-going educational systems. Concerning the results of the institution B, it was reported that the majority of the participants (N= 14; P= 73.7%) confirmed taking the overall evaluation of the program and current education systems into consideration while carrying out the testing procedures. On the other hand, 26.3% (N= 5) of the participants were not sure about it. With a mean score of 3.79/ 5.00 (SD= .54), the participants from the institution B held similar opinion with each other. Hence, it could be stipulated that participants from the institution B dominantly accepted the presence of a validation of the testing procedures at the helm of the overall evaluation of the total program together with the current educational systems in use.

Similarly, it was secondarily checked whether any standards were set before administering the tests in order to disseminate quality in related testing and assessment practices. In relation with this, 84.2% (N= 16) of the participants confirmed it. Yet, the rest as the 15.8% (N= 3) of them was not sure whether that was the case in practice. With a mean score of 4.11/ 5.00 (SD= .66), the participants from the institution B held similar opinion with each other. Therefore, it could be stipulated that some core elements were defined as the standards of testing and assessment practices before the tests were administered to the target population. Similarly, it was also probed whether the tests in use covered a wide range of cultural and educational contexts within. Herein, it was reported by the majority of the participants (N= 17; P= 89.5%) that the tests in use were constituted by various cultural and educational contexts within. However, 10.5% (N= 2) of them was not sure whether variation in contextual usage of the target language was addressed by

the tests in use. With a mean score of 4.21/ 5.00 (SD= .66), the participants from the institution B held similar opinion with each other.

With a view to the guiding principles ascertained by the AEA- Europe, it was checked whether a European perspective was pursued while designing tests, and developing test items. Herein, 89.5% (N= 17) of the participants from the institution B confirmed adopting a European perspective in designing tests. However, 10.5% (N= 2) of the participants were not sure whether that was the case. With a mean score of 4.11/ 5.00 (SD= .57), the participants from the institution B were of similar opinion with each other. Correlatively, it was scrutinized whether innovative assessment techniques were used by the private institutions previously selected. In this context, 84.3% (N= 16) of the participants confirmed it. However, 15.8% (N= 3) of the participants were not sure whether innovative techniques were adopted in testing and assessment practices. With a mean score of 4.05/ 5.00 (SD= .62), the participants from the institution B held similar opinion with each other.

Additionally, it was also checked whether the assessment process did run taking the place of the test takers within the testing procedure into consideration. Herein, 78.9% (N= 15) of the participants from the institution B asserted that the assessment process did not run independently, instead took the place of the test takers within the testing procedure into consideration. Yet, 21.1% (N= 4) of them were not sure about it. With a mean score of 4.16/ 5.00 (SD= .76), the participants from the institution B were of similar opinion with each other. Concomitantly, it was questioned whether the test takers' rights were concerned in the assessment process. Herein, 84.2% (N= 16) of the participants confirmed it. But 15.8% (N= 3) of them were not sure whether they were the rights of the ones who did devise and administer the tests which were protected, albeit not the ones who took the tests as the candidates. With a mean score of 4.11/ 5.00 (SD= .66), the participants from the institution B were of similar opinion with each other.

Relatively, it was checked whether the alignment of the individual's place in the assessment procedure with the United Nations Convention on the Rights of the Child was available. In this vein, the majority of the participants (N= 15; P= 79%) from the institution B was not sure whether that was the case. On the other hand, 21.1% (N= 4) of them confirmed it. With a mean score of 4.00/ 5.00 (SD= .67), the participants from the institution B held similar opinion with each other. Therefore, it

could be stipulated that the majority of the participants from the institution B was not well aware of the presence of such an alignment between the test takers' rights and the United Nations Convention on the Rights of the Child. To note more, it was scrutinized whether ethical issues were taken into consideration within assessment procedures. Herein, 73,7% (N= 14) of the participants from the institution B confirmed it. Yet, 26.3% (N= 5) of them were not sure whether that was the case. With a mean score of 3.84/ 5.00 (SD= .60), the participants from the institution B held similar opinion with each other.

With a view to the guiding principles ascertained by the AEA- Europe, it was checked whether the essentials of assessment procedure were marked properly. In this case, 89.5% (N= 17) of the participants from the institution B confirmed that the keystones of assessment were properly addressed. On the other hand, 10.5% (N= 2) of them were not sure about it. With a mean score of 3.95/ 5.00 (SD= .40), the participants from the institution B held similar opinion with each other. Additionally, it was questioned whether the test results were functional and purposeful. Herein, almost all of the participants (N= 18; P= 94.7%) from the institution B confirmed it. Yet, 5.3% (N= 1) of them were not sure about it. With a mean score of 4.32/ 5.00 (SD= .58), the participants from the institution B were of very similar opinion with each other. Correlatively, it was checked whether the test results could be utilized in other educational settings for further use. Herein, 89.5% (N= 17) of the participants from the institution B confirmed it in contrast with the 10.5% (N= 2) of them, stating that they were not sure whether the test results could be used afterwards for any other educational purpose. With a mean score of 4.21/ 5.00 (SD= .63), the participants from the institution B held similar opinion with each other.

With a view to the guiding principles ascertained by the AEA- Europe, it was also scrutinized whether the overall education of the test takers was supported by the tests in use. In this context, all of the participants from the institution B (N= 19; P= 100%) confirmed it. With a mean score of 4.16/ 5.00 (SD= .37), the participants from the institution B were of the same opinion with each other. Concomitantly, it was delved whether that the test takers were provided with tests which could assign its rationale on the intended learning in tow. Herein, the majority of the participants from the institution B (N= 16; P= 84.2%) confirmed it whereas 15.8% (N= 3) of them

were not sure about it. With a mean score of 4.00/ 5.00 (SD= .58), the participants from the institution B held similar opinion with each other.

In the same vein, it was checked whether the tests in use were correlated with the rudiments of the CEFR. Herein, 89.4% (N= 17) of the participants from the institution B confirmed it. However, 10.5% (N= 2) of them were not sure whether the tests in use provided information on the adoption of the objectives set by the CEFR. With a mean score of 4.00/ 5.00 (SD= .47), the participants from the institution B held similar opinion with each other. Additionally, it was questioned whether the basic tenets of the CEFR were pursued by the test development cycle. Accordingly, more than half of the participants (N= 15; P= 79%) confirmed it. On the other hand, 21.1% (N= 4) of them were not sure whether that was the case. With a mean score of 4.00/ 5.00 (SD= .67), the participants from the institution B held similar opinion with each other.

In respect to the guiding principles ascertained by the AEA- Europe, it was also scrutinized whether there were different types of assessment in use which enabled getting feedback on the on-going educational system. Accordingly, the majority of the participants (N= 17; P= 89.5%) from the institution B confirmed it in spite of the 10.5% (N= 2) of them who were not sure about it. With a mean score of 4.21/ 5.00 (SD= .63), the participants from the institution B held similar opinion with each other. Correlatively, it was checked whether different assessment types in use together with the test results gained allowed for the evaluation of the program by decision-makers. Herein, 89.5% (N= 17) of the participants from the institution B confirmed that the decision-makers could have the opportunity to evaluate the current program by means of the test results gained. Yet, 10.5% (N= 2) of them were not sure whether that was the case. With a mean score of 4.16/ 5.00 (SD= .60), the participants from the institution B held similar opinion with each other. Additionally, it was probed whether potential evidences were given after the implementation of each test in order to check whether the standard requirements previously defined were met, or not. Herein, 84.2% (N= 16) of the participants confirmed it whereas 15.8% (N= 3) of them were not sure about it. With a mean score of 3.95/ 5.00 (SD= .52), the participants from the institution B held similar opinion with each other.

With special reference to the instrument ascertained by the AEA- Europe, the assessment types in use as an instrument for testing and assessment practices were questioned. Accordingly, almost all of the participants (N= 18; P= 94.7%) from the institution B confirmed the exploitation of standardized tests within. Besides, 9.1% (N= 1) of them disagreed that standardized tests were conducted within the institution B. With a mean score of 4.32/ 5.00 (SD= .58), the participants from the institution B held very similar opinion with each other. At the very same, it was certified by the majority of the participants (N= 14; P= 73.7%) that summative assessment such as school-based examinations were applied within the institution B. Yet, 26.3% (N= 5) of them were not sure whether such an assessment type was being conducted. With a mean score of 3.95/ 5.00 (SD= .71), the participants from the institution B held similar opinion with each other. Apart from these, it was asserted by the majority of the participants (N= 14; P= 73.7%) that performance assessment was conducted as an implementation within the institution B. However, the rest (N= 5; P= 26.3%) was not sure whether that was the case. With a mean score of 3.95/ 5.00 (SD= .71), the participants from the institution B held similar opinion with each other.

To add more, almost all of the participants (N= 18; P= 94.7%) from the institution B confirmed that formative assessment was conducted within. Yet, 5.3% (N= 1) of them were not sure whether learning outcomes of a curriculum were covered by means of formative assessment. With a mean score of 4.05/ 5.00 (SD= .40), the participants from the institution B held very similar opinions with each other. Likewise, the majority of the participants (N= 16; P= 84.2%) from the institution B confirmed that competency tests were in use within the institution B. Yet, 15.8% (N= 3) of them were not sure whether that was the case. With a mean score of 4.00/ 5.00 (SD= .58), the participants from the institution B held similar opinion with each other.

The implementation of the AEA- Europe's framework of standards by the institution C. An overall estimation regarding the results gained from all of the private institutions were checked, and reported together and separately. With this in mind, the overall results of the implementation of the Framework of Standards by the AEA-Europe are listed below regarding the case in private institution C.

With a view to the guiding principles ascertained by the AEA- Europe, it was initially checked whether the testing procedures were framed by the overall evaluation of the program and on-going educational systems. Concerning the results of the institution C, it was reported that the majority of the participants (N= 8; P= 80%) confirmed taking the overall evaluation of the program and current education systems into consideration while carrying out the testing procedures. On the other hand, 10% (N= 1) of the participants were not sure about it. Besides, 10% (N= 1) of them disagreed with them, though. With a mean score of 4.00/ 5.00 (SD= .94), the participants from the institution C held similar opinion with each other. Hence, it could be stipulated that participants from the institution C predominantly accepted the presence of a validation of the testing procedures at the helm of the overall evaluation of the total program together with the current educational systems in use.

With a view to the guiding principles ascertained by the AEA- Europe, it was secondarily checked whether any standards were set before administering the tests in order to disseminate quality in related testing and assessment practices. In relation with this, 60% (N= 6) of the participants confirmed it. Yet, the rest as the 40% (N= 4) of them were not sure whether that was the case in practice. With a mean score of 3.60/ 5.00 (SD= .52), the participants from the institution C held different opinions from each other. Therefore, it could be stipulated that some core elements were assumed by the majority of the participants from the institution C to be defined as the standards of testing and assessment practices before the tests were administered to the target population.

Similarly, it was checked whether a European perspective was pursued while designing tests, and developing test items. Herein, 60% (N= 6) of the participants from the institution C confirmed adopting a European perspective in designing tests. However, 30% (N= 3) of the participants were not sure whether that was the case. Besides, 10% (N= 1) of them disagreed with them, though. With a mean score of 3.50/ 5.00 (SD= .71), the participants from the institution C were of similar opinion with each other. Correlatively, it was scrutinized whether innovative assessment techniques were used by the private institutions previously selected. In this context, 70% (N= 7) of the participants were not sure whether innovative techniques were adopted in testing and assessment practices. However, 30% (N= 3) of the

participants confirmed it. With a mean score of 3.30/ 5.00 (SD= .48), the participants from the institution C held similar opinion with each other. Henceforth, it could be stipulated that the majority of the participants from the institution C was not well aware of the novelty of the assessment techniques in use within the institution C.

With a view to the guiding principles ascertained by the AEA- Europe, it was also probed whether the tests in use covered a wide range of cultural and educational contexts within. Herein, it was reported by the half of the participants (N= 5; P= 50%) that the tests in use were constituted by various cultural and educational contexts within. However, 40% (N= 4) of them were not sure whether variation in contextual usage of the target language was addressed by the tests in use. Besides, 10% (N= 1) of them dissented to it, though. With a mean score of 3.40/ 5.00 (SD= .70), the participants from the institution C held different opinions from each other.

In the same vein, it was also checked whether the assessment process did run taking the place of the test takers within the testing procedure into consideration. Herein, 60% (N= 6) of the participants from the institution C asserted that the assessment process did not run independently, instead took the place of the test takers within the testing procedure into consideration. Yet, 40% (N= 4) of them were not sure about it. With a mean score of 3.70/ 5.00 (SD= .67), the participants from the institution C held different opinions from each other. Concomitantly, it was questioned whether the test takers' rights were concerned in the assessment process. Herein, there were contradictory arguments between the ones confirming it (N= 4; P= 40%) and the ones who were not sure whether they were the rights of the ones who did devise and administer the tests which were protected, albeit not the ones who took the tests as the candidates (N=4; P= 40%). Additionally, 20% (N= 2) of them disagreed with them, though. With a mean score of 3.20/ 5.00 (SD= .79), the participants from the institution C were of different opinions from each other.

Relatively, it was checked whether the alignment of the individual's place in the assessment procedure with the United Nations Convention on the Rights of the Child was available. In this vein, the majority of the participants (N= 6; P= 60%) from the institution C was not sure whether that was the case. On the other hand, 30% (N= 3) of them confirmed it. Besides, 10% (N= 1) of them disagreed with them,

though. With a mean score of 3.20/ 5.00 (SD= .63), the participants from the institution C held similar opinion with each other. Therefore, it could be stipulated that the majority of the participants from the institution C was not well aware of the presence of such an alignment between the test takers' rights and the United Nations Convention on the Rights of the Child. To note more, it was scrutinized whether ethical issues were taken into consideration within assessment procedures. Herein, 60% (N= 6) of the participants from the institution C confirmed it. Yet, 40% (N= 4) of them were not sure whether that was the case. With a mean score of 3.60/ 5.00 (SD= .52), the participants from the institution C held different opinions from each other.

With a view to guiding principles ascertained by the AEA- Europe, it was checked whether the essentials of assessment procedure were marked properly. In this case, half of the participants (N= 5; P= 50%) from the institution C was not sure whether the keystones of assessment were properly addressed. On the other hand, the other half of them (N= 5; P= 50%) confirmed it. With a mean score of 3.60/ 5.00 (SD= .70), the participants from the institution C held different opinions from each other. Additionally, it was questioned whether the test results were functional and purposeful. Herein, almost all of the participants (N= 8; P= 80%) from the institution C confirmed it. Yet, 20% (N= 2) of them were not sure about it. With a mean score of 3.80/ 5.00 (SD= .42), the participants from the institution C were of similar opinion with each other.

Correlatively, it was checked whether the test results could be utilized in other educational settings for further use. Herein, half of the participants (N= 5; P= 50%) from the institution C confirmed it in contrast with the 30% (N= 3) of them, stating that they were not sure whether the test results could be used afterwards for any other educational purpose. Besides, 20% (N= 2) of them strongly disagreed with them, though. With a mean score of 3.20/ 5.00 (SD= 1.32), the participants from the institution C held different opinions from each other. Additionally, it was also scrutinized whether the overall education of the test takers was supported by the tests in use. In this context, the majority of the participants from the institution C (N= 8; P= 80%) confirmed it. Yet, 20% (N= 2) of the participants from the institution C were not sure about it. With a mean score of 3.80/ 5.00 (SD= .42), the participants from the institution C were of similar opinion with each other. Concomitantly, it was

delved whether that the test takers were provided with tests which could assign its rationale on the intended learning in tow. Herein, the majority of the participants from the institution C (N= 8; P= 80%) confirmed it whereas 20% (N= 2) of them were not sure about it. With a mean score of 3.80/ 5.00 (SD= .42), the participants from the institution C held similar opinion with each other.

With reference to the guiding principles ascertained by the AEA- Europe, it was checked whether the tests in use were correlated with the rudiments of the CEFR. Herein, 70% (N= 7) of the participants from the institution C confirmed it. However, 30% (N= 3) of them were not sure whether the tests in use provided information on the adoption of the objectives set by the CEFR. With a mean score of 3.70/ 5.00 (SD= .48), the participants from the institution C held similar opinion with each other. Additionally, it was questioned whether the basic tenets of the CEFR were pursued by the test development cycle. Accordingly, there was a contradiction between the ones who confirmed it (N= 4; P= 40%) and the ones who were not sure about it (N= 4; P= 40%). On the other hand, 20% (N= 2) of them disagreed with it, though. With a mean score of 3.20/ 5.00 (SD= .79), the participants from the institution C held different opinions from each other.

Besides, it was also scrutinized whether there were different types of assessment in use which enabled getting feedback on the on-going educational system. Accordingly, more than half of the participants (N= 6; P= 60%) from the institution C confirmed it in spite of the 40% (N= 4) of them who were not sure about it. With a mean score of 3.70/ 5.00 (SD= .67), the participants from the institution C held different opinion from each other. Correlatively, it was checked whether different assessment types in use together with the test results gained allowed for the evaluation of the program by decision-makers. Herein, 80% (N= 8) of the participants from the institution C confirmed that the decision-makers could have the opportunity to evaluate the current program by means of the test results gained. Yet, 20% (N= 2) of them were not sure whether that was the case. With a mean score of 4.10/ 5.00 (SD= .74), the participants from the institution C held similar opinion with each other. Additionally, it was probed whether potential evidences were given after the implementation of each test in order to check whether the standard requirements previously defined were met, or not. Herein, 60% (N= 6) of the participants were not sure about it whereas 30% (N= 3) of them confirmed it. With a mean score of 3.20/

5.00 (SD= .63), the participants from the institution C held similar opinion with each other.

With special reference to the instrument ascertained by the AEA- Europe, the assessment types in use as an instrument for testing and assessment practices were questioned. Accordingly, more than half of the participants (N= 6; P= 60%) from the institution C confirmed the exploitation of standardized tests within. Besides, 30% (N= 3) of them were not sure about it together with the other 10% (N= 1) who were strongly disagreed that standardized tests were conducted within the institution C. With a mean score of 3.40/ 5.00 (SD= .97), the participants from the institution C held similar opinion with each other. At the very same, it was certified by the half of the participants (N= 5; P= 50%) that summative assessment such as school-based examinations were applied within the institution C. Yet, 40% (N= 4) of them were not sure whether such an assessment type was being conducted. Besides, 10% (N= 1) of them strongly disagreed with it, though. With a mean score of 3.30/ 5.00 (SD= .95), the participants from the institution C held different opinions from each other.

Apart from these, it was asserted by the majority of the participants (N= 7; P= 70%) that performance assessment was conducted as an implementation within the institution C. However, 20% (N= 2) of the participants from the institution C were not sure whether that was the case. Besides, 10% (N= 1) of them strongly disagreed with it, though. With a mean score of 3.50/ 5.00 (SD= .97), the participants from the institution C held similar opinion with each other. To add more, there were contrary arguments between the ones who asserted that formative assessment was conducted within the institution C (N= 4; P= 40%) and the ones who were not sure whether that was the case within (N= 4; P= 40%). Besides, 20% (N= 2) of them asserted that learning outcomes of a curriculum were not covered by means of formative assessment. With a mean score of 3.20/ 5.00 (SD= .79), the participants from the institution C held different opinions from each other. Likewise, the majority of the participants (N= 8; P= 80%) from the institution C confirmed that competency tests were in use within the institution C. Yet, 10% (N= 1) of them were not sure whether that was the case. Besides, 10% (N= 1) of them disagreed with it, though. With a mean score of 3.70/ 5.00 (SD= .67), the participants from the institution C held similar opinion with each other.

The overall picture of the implementation of the AEA- Europe's framework of standards by selected private institutions. The Framework of Standards by the AEA- Europe were summed up in two basic components within the questionnaire used for this study. These components were the 'guiding principles' and 'instrument'. Composed of 24 test items in total, these two sub-sections were analyzed separately, and the estimations gained were reported singly.

Accordingly, the highest number of participants (N= 36) were estimated to either 'agree' or 'strongly agree' on the component of the guiding principles, who confirmed at the ratio of 90% that the tests results were meaningful and purposeful. Hence, it could be inferred that the test results were utilizable in other settings for further educational purposes. It was followed by the assumption that the overall education of the test takers was supported by the purpose of the assessment at the ratio of 87.5%. Therefore, it could be stipulated that the majority of the participants (N= 35) was convinced of the reciprocal relationship between the purpose of the assessment and the overall education of the test takers by the institution they were working at. Likewise, it was postulated by the majority of the participants (N= 34; P= 85%) that the test results made it possible for decision makers to evaluate the on-going programs and designate resources. Therefore, it could be stipulated that the test results could be used for further cases by decision makers for the evaluation of the current program.

By the same token, it was asserted by the majority of the participants (N= 33; P= 82.5%) that the intended learning was screened by the assessment implementations conducted within the previously selected private institutions. Correspondingly, the same ratio (N= 33; P= 82.5%) of participants alleged that the assessment types provided feedback on the educational system in practice. In conducting assessment practices within their bodies, the previously selected private institutions were stated to cover competency tests in the main (N= 33; P= 82.5%). Henceforth, it could be stipulated that the intended learning was most generally assessed by competency tests, the results of which catered feedback for the current educational system. Additionally, it was asserted by the majority of the participants (N= 32; P= 80%) that the overall evaluation of the total program together with the review of the current educational system were considered while conducting testing

and assessment practices. In addition to this, the majority of the participants (N= 32; P= 80%) asserted that the assessment procedures followed within the previously selected private institutions were in rapport with the objectives set by the CEFR. In the light of these, it could be stipulated that the assessment procedures were conformed with the aims of the CEFR, which made contribution to the overall evaluation of the total program and on-going educational system.

Besides, when it was questioned whether on-going testing and assessment practices were molded with some established set of standards in order to disseminate quality in relevant practices, it blossomed as a result that 77.5% (N= 31) of the participants confirmed it. In the same vein, the same ratio (N= 31; P= 77.5%) of participants asserted that the assessment results could be permissible in other educational settings, as well. To note more, the majority of the participants (N= 31; P= 77.5%) stated that the assessment procedures conducted within the previously selected private institutions were trimmed by the learning outcomes of a curriculum, which was basically labelled as formative assessment.

One more to note, the majority of the participants (N= 30; P= 75%) confirmed that the tests in use were embellished with various types of educational and cultural contexts. Herein, 22.5% (N= 9) of them were not sure whether that was the case. Following that, the majority of the participants (N= 29; P= 72.5%) asserted that the keystones of assessment such as reliability, validity and practicality were cautiously handled. Herein, 22.5% (N= 9) of them were not sure whether that was the case. In the same vein, the majority of the participants (N= 29; P= 72.5%) confirmed that the assessment applied within the selected private institutions was merged in standardized tests to some extent. On the other hand, 22.5% (N= 9) of them were not sure whether that was the case. In this context, the participants who felt indecisive with regard to the type of tests in use could not be underestimated as the proportion of the indecisive participants was barely one-fourth of the number of participants in total.

Besides, although the number of participants (N= 28; P= 70%) who confirmed that the assessment applied within the previously selected private institutions was conducted by summative tests was higher than those who were not sure about it (N= 11; P= 27.5%) together with the number of participants who disagreed (N= 1; P= 2.5%), the amount of the followings in tow could not be disregarded as there

were 40 participants in total. In the same vein, the same ratio of participants (N= 28; P= 70%) confirmed that the assessment procedures conducted within those selected private institutions were framed by a pre-defined European perspective of world-wide interest. However, there were 11 of them (P= 27.5%) who were not sure whether that was the case, though. Additionally, the majority of the participants (N= 28; P= 70%) asserted that the test takers' places in the assessment procedure were accurately pinpointed whereas there were 12 of them (P= 30%) who were not sure whether that was the case. Therefore, it could be speculated that not all of the participants in this study, the English language teachers who were also working as test (-item) developers at those private institutions, were well-aware of the types of assessment conducted within.

At the very same, they were not well-aware of the adoption of a European perspective in assessment procedures along with the sum of participants who were not apprised of the test takers' places in the assessment procedure. In addition to these, the majority of the participants (N= 27; P= 67.5%) confirmed that innovative assessment techniques were scrutinized in the course of test preparation and (test-) item design. Nevertheless, the number of participants who were not sure about it (N= 13; P= 32.5%) could not be underestimated as it was slightly higher than the one-fourth of the participants in sum who were totally 40 in number. Therefore, it could be stipulated that the test (-item) design procedure did not thoroughly cover the adoption of innovative techniques in testing and assessment. Correlatively, the same ratio of participants (N= 27; P= 67.5%) confirmed that ethical issues were considered while conducting assessment procedures. Herein, more than one-fourth of the participants (N= 13; P= 32.5%) was not sure whether that was the case.

At the very same, the same ratio of participants (N= 27; P= 67.5%) asserted that assessment procedures were framed by the rights of the test takers rather than the test (-item) developers and administrators. On the other hand, more than one-fourth of the participants (N= 11; P= 27.5%) was not sure whether test takers' rights in the assessment procedure were attached more importance than those of designers and administrators. Along the same line, the majority of the participants (N= 27; P= 67.5%) approved that the assessment applied in the selected private institutions was laced with performance assessment. However, more than one-fourth of the participants (N= 12; P= 30%) was not sure whether that was the case.

The number of participants who were not sure of those four items mentioned above could not be underestimated as it was slightly higher than one-fourth of the participants in sum who were totally 40 in number. Therefore, it could be stipulated that not all of the participants of this study were aware of the on-going assessment procedures in their institutions thoroughly.

Additionally, when it was questioned whether any possible evidences were presented after the implementation of each test to check the appropriateness of tests to the standard requirements predetermined, it blossomed as a result that 65% (N= 26) of the participants either agreed or strongly agreed. However, there were 13 of them (P= 32.5%) who were not sure about it, though. Hence, it could be indicated that slightly more than one-fourth of the participants did not rest assured of the alignment of the tests to the standards set before the implementation of each test. In the same vein, the majority of the participants (N= 26; P= 65%) confirmed that the assessment procedure implemented within the previously selected private institutions followed the test development cycle purported by the CEFR. Herein, 30% (N= 12) of the participants was not sure whether that was the case. From this point forth, it could be stipulated that slightly more than one-fourth of the participants was not sure whether the tests conducted within those private institutions were aligned with the CEFR as well as being appropriate to the previously set standards.

One more to note, the fewest number of participants (N= 22; P= 55.5%) confirmed that the assessment procedure followed within the previously selected private institutions aligned with the United Nations Convention on the Rights of the Child. On the other hand, the rest as the other majority was either not sure about (N= 17; P= 42.5%), or disagreed with it (N= 1; P= 2.5%). In the light of this, it could be stipulated that nearly half of the participants from those selected private institutions was not well-aware of what the aforementioned rights were basically about.

What is the general paradigm of a sample of leading professionals from selected non-formal English language schools in Turkey (i.e. decision-makers, testing office, English language teachers) on the implementation of testing and assessment procedures as defined by the European guidelines?

The congruence of the testing and assessment practices with the above mentioned European guidelines was pinpointed by five questions answered and reported

separately. Together with this, the overall picture of each private institution was framed by the European standards, namely the CEFR, ALTE, ILTA and AEA-Europe. For each of these European standards, the components were defined, labelled and discussed with the estimations gathered with great extent of scope. Thus, the procedural outlines of the previously selected private institutions' implementations of testing and assessment were marked.

In addition to these, the general paradigm of a sample of leading professionals from a range of non-formal English language schools in Turkey on the implementation of testing and assessment procedures as defined by the European guidelines was drawn taking the views of the decision-makers and English language teachers, who were also working as test (-item) developers at the same institutions. Herein, the results were presented in a two-way alternate. Firstly, the overall estimations regarding the exploitation of all European standards by selected private institutions was reported by means, standard deviations and standard errors of mean for each of them elaborately. In this context, the replies of the English language teachers to the questionnaire were noted at one hand. Secondly, the viewpoints of the directors from the selected private institutions were addressed by their own answers gathered from the semi-structured interview sessions to the accompaniment of 6 questions, which were listed below in a detailed way. Additionally, the viewpoints of the director of ÖZ-KUR-DER were also noted in order to frame the outline better, and to enable a triangulation by the answers gathered from the English language teachers, the directors of the selected private institutions and the director of ÖZ-KUR-DER.

In the light of these, this part was constituted by two sub-sections, in which the utilization of the European guidelines in testing and assessment practices by selected private institutions, and the viewpoints of the directors from those private institutions and ÖZ-KUR-DER on the utilization of the European guidelines in testing and assessment practices were reported separately in order to draw a general paradigm for the implementation of the aforementioned European guidelines in testing and assessment practices by the previously selected private institutions.

The utilization of the European guidelines in testing and assessment practices by selected private institutions. The general paradigm of a sample of leading professionals from a number of non-formal English language schools in

Turkey on the implementation of testing and assessment procedures as defined by the European guidelines was drawn taking the views of the decision-makers and English language teachers, who were also working as test (-item) developers at the same institutions in a two-way alternate. Accordingly, as the first step, the overall estimations regarding the exploitation of all European standards by selected private institutions was reported by means, standard deviations and standard errors of mean for each of them elaborately. In this context, the replies of the English language teachers to the questionnaire were noted at one hand, and the results of each private institution were reported separately, and the results were presented in the table given below:

Table 19

The Utilization of the European Guidelines in Testing and Assessment Practices by Selected Private Institutions

The European Standard(s)	Private Institution(s)	Mean	Std. Error of Mean	Std. Deviation
ALTE	A	3.75	.116	.384
	B	4.01	.065	.284
	C	3.47	.157	.496
ILTA	A	4.13	.128	.424
	B	4.00	.095	.414
	C	3.97	.136	.429
EALTA	A	3.67	.089	.294
	B	3.97	.074	.320
	C	3.55	.158	.499
AEA- Europe	A	3.84	.096	.317
	B	4.07	.067	.291
	C	3.53	.112	.355
Sum	A	3.85	.085	.283
	B	4.01	.060	.262
	C	3.63	.117	.369

The overall results above showed that the highest mean score of all selected private institutions regarding the utilization of the ALTE Code of Practice was estimated by the private institution B (M= 4.01; SD= .06). It was followed by the institution A (M= 3.75; SD= .12) and institution C (M= 3.47; SD= .16) respectively. In a similar vein, the highest mean score of all selected private institutions regarding the utilization of

the EALTA Guidelines was estimated by the private institution B (M= 3.97; SD= .07), which was followed by that of institution A (M= 3.67; SD= .09) and that of institution C (M= 3.55; SD= .16) respectively. Within the scope of the utilization of the ILTA Guidelines for Practice, the highest mean score of all was estimated by the private institution A (M= 4.13; SD= .13), which was followed by that of institution B (M= 4.00; SD= .09) and that of institution C (M= 3.97; SD= .14) respectively. For the utilization of the Guidelines of the AEA- Europe, the highest mean score was estimated by the private institution B (M= 4.07; SD= .07), which was followed by that of institution A (M= 3.84; SD= .10) and that of institution C (M= 3.53; SD= .11) respectively.

In the light of these, the private institution B was reported as applying the above-mentioned European guidelines more than the others with the highest mean score of all (M= 4.01; SD= .06). It was pursued by the private institution A with the mean score of 3.85/ 5.00 (SD= .08). On the other hand, the lowest mean score of all was estimated by the private institution C (M= 3.63; SD= .12). Therefore, it could be stipulated that the private institution B outscored the other private institutions regarding the utilization of some European guidelines in language testing and assessment practices.

The viewpoints of the directors from the selected private institutions and ÖZ-KUR-DER on the utilization of the European guidelines in testing and assessment practices. The viewpoints of the directors from the selected private institutions on the utilization of the European guidelines in testing and assessment practices were highlighted by means of semi-structured interview sessions conducted face-to-face. Herein, 6 questions were addressed by the researcher to get more detailed information on the gist of the testing and assessment practices conducted within the institutions and their alignment to the pre-defined European guidelines. Accordingly, these questions were listed as:

1. Please provide some information on the testing and assessment practices conducted within your institution(s).
2. Are these practices aligned with any European standards? If yes, please provide some information about those standards.
3. Please provide some information about the instruments and the criteria set for testing and assessment practices.

4. Please provide some information on the difficulties and problems mostly encountered in conducting testing and assessment practices.
5. Please provide some recommendations in order to enhance the on-going testing and assessment practices within your institution(s).
6. Please provide some recommendations in order to enhance the on-going testing and assessment practices across the country.

In the light of these, each of the private institution was singly probed within the scope of above mentioned 6 questions. The answers were noted pursuant to the directors' standpoints on the current implementations in testing and assessment. In doing this, for each private institution, the directors' answers on the above listed 6 questions were elaborated in detail encompassing information on the testing and assessment practices conducted within the private institution(s), information on the alignment of the testing and assessment practices with the European guidelines, information on the instruments in use and criteria set for testing and assessment practices, information on the difficulties and problems mostly encountered in conducting testing and assessment practices, together with the recommendations on the development of the current testing and assessment practices within the private institution(s) and across the country. Therefore, the differences in testing and assessment practices among the selected private institutions were also detected. Not to mention, the viewpoints of the director of ÖZ-KUR-DER were also reported below in order to frame the outline better, constituting one of the wings of the aforementioned triangulation.

The directors' viewpoints from the institution A. Each of the directors was initially asked to give some information on the testing and assessment practices conducted within their institutions. Accordingly, the director of the private institution A stated that all of the students enrolled solely with an age of either 15 and above were using the ELP as a tool for self-assessment. It was also added by the director of A that there might be some other applications (e.g. pop quizzes) conducted by the English language teachers, as well. As the ELP was in use, and the classes were defined by the proficiency levels from A1 to C2 although A2 was marked as 'elementary', it could be stipulated that the assessment was aligned with the CEFR at some points.

To elaborate the examinations in use, the director of A asserted that after 80 hours of lecture (as each of the classes lasted for 80 hours in total within the institution A), the students were taken to a placement test. The placement tests were comprised of test items on vocabulary, grammar, listening and reading, which were all prepared by the testing office. On the other hand, testing and assessing of speaking and writing was left to the English language teachers of those classes. Thus, it could be stipulated that there was an independence on the testing and assessment of productive language skills in terms of English language teacher, class and day.

With respect to the difficulties and problems mostly encountered in conducting testing and assessment practices, it was reported by the director of A that the private institution(s) appeared as a trading house which was merchandizing education. Therefore, the student(s) enrolled in such kind of private institution(s) were well aware of the fact that it was the identity of the institution(s) which was protected, albeit not that of student(s). To set an example, the director of A stated that if there was a vacancy in A2-level proficiency class, a student who was marked as proficient at B1 level via placement test was also sent to that class due to the fact that B1-level proficiency class was full. Moreover, the tests were conducted in multiple-choice-item format within the scope of vocabulary, grammar, listening and reading. Besides, each English language teacher prepared his/her own speaking and writing examinations, and conducted these examinations at his/her convenience. Thereafter, a mean value was calculated to get a final score for the placement test. As there was no standards in testing and assessment of speaking and writing, the director of A reported that some a priori problems might mushroom as a result of misapplications.

In other respects, for the enhancement of on-going testing and assessment practices within the institution A, its director recommended that performance assessment was to be placed more importance than paper-and-pencil tests. Postulated as the fundamentals of language teaching by the director of A, the productive skills were suggested to be given more prominence by even creating and adopting a new form of placement test based on an oral proficiency examination, as well. For the improvement of the on-going testing and assessment practices across the country, the director of A stated that a skills-based approach was to be employed

by all education centers; henceforth, the students enrolled in any of those centers could internalize the English language better.

The directors' viewpoints from the institution B. Each of the directors initially asked to give some information on the testing and assessment practices conducted within their institutions. Accordingly, the director of the private institution B stated that it was a must for all of the students enrolled in the institution to use the ELP as a tool for self-assessment. It was also added by the director of B that there were some other applications in use, such as pop quizzes for each of the language skills separately, as well. As the ELP was in use, and the classes were defined by the proficiency levels from A1 to C2, it could be stipulated that the assessment was aligned with the CEFR.

To elaborate the examinations in use, the director of B asserted that the students enrolled in the institution B were taken to a diagnostic test in order to determine the level of language proficiency at the outset. Particularly, this diagnostic test was done on students' speaking skill, and the results gathered made it possible to know where the students were academically so as to bring them to where they were actually in need to be. Correlatively, after 80 hours of lecture (as each of the classes lasted for 80 hours in total within the institution B), the students were taken to a placement test to spot the proficiency levels of each. The placement test was also done on the speaking skill. Therefore, it could be stipulated that a skills-based assessment was adopted, and there were no paper-and-pencil tests in use, albeit oral proficiency examinations instead. One more to note, these oral proficiency examinations were evaluated by two independent raters, of one whom was a native speaker of the target language.

Additionally, it was reported by the director of B that there were various student clubs on a language-skill-basis such as vocabulary club, speaking club, grammar club so that the students could have the opportunity to experience the authentic use of the target language. Every student was to be enrolled for at least one club, and to join the activities organized within. Besides, the director of B stated that there were some other optional career clubs, such as how-to-prepare-a-CV club, how-to-get-prepared-for-an-interview-in-English club and the like. Likewise, it was reported by the director of B that all of the English language teachers working in the private institution B were expected to get in-service training either on-the-job

or by the head office. These in-service training activities could include the effective use of body language, the art of rhetoric, teaching English to speakers of other languages (TESOL) and the like. Unlike the private institution A, where a student was placed in a A2-level class just because there was a vacancy although s/he was in fact proficient at the level of B1, a student was placed at a proficiency class according to the results of an oral proficiency examinations conducted. Besides, if a student quit the course when s/he was at B1 proficiency level and came back some time later, then that student was to start the course from the very beginning; thereby, that student was placed at an A1-level class.

With respect to the recommendations for the enhancement of on-going testing and assessment practices within the institution B, its director stated that the students were to be given freedom so that they could quiet their minds, and feel free to speak when they did feel truly ready. For the improvement of the on-going testing and assessment practices across the country, the director of the private institution B recommended that Turkish system of English language teaching led by the MoNE was to be revised and modernized so as not to be out-of-date. To set an example for this, the director of B addressed that English language teaching could be a part of early childhood education and/or pre-school education, and be a prerequisite for further education. In the same context, it was marked out by the director of B that the ELT curriculum was to be reviewed as the newly graduates of the ELT departments in Turkey had some problems in conducting skills-based testing and assessment procedures. To add more, the director of B suggested that there was to be a standardization in testing and assessment practices across the country. Because someone with a proficiency level of B1 might be regarded as proficient at the level of A2 by another institution.

The directors' viewpoints from the institution C. Each of the directors was initially asked to give some information on the testing and assessment practices conducted within their institutions. Accordingly, the director of the private institution C stated that there were 60 students in total, who were using the ELP as a tool for self-assessment within the institution C. More generally, it was noted by the director of C that there were approximately 1.000 students in sum using the ELP in other branches of the private institution C across the country. Herein, it was also noted by its director that the ELP was used by the students who were either elementary level

or above. Besides, the director of C stated that the CEFR was adopted in testing and assessment practices. As the ELP was in use, and the classes were defined by the proficiency levels from A1 to C2 although A2 was marked as 'elementary', it could be stipulated that the assessment was aligned with the CEFR at some points.

To elaborate the examinations in use, the director of C asserted that 4 midterms, 2 progress and 1 oral proficiency examinations were conducted for each of the classes. To set an example for this, a student at B1 proficiency level was to take all of the examinations above listed, and to be successful to get through to the proficiency level of B2. It was also added by the director of C that there might be some other applications (e.g. pop quizzes) conducted by the English language teachers, as well.

With respect to the difficulties and problems mostly encountered in conducting testing and assessment practices, it was reported by the director of C that the most difficult part was the teachers' internalization of the new applications as it was marked as rather hard to persuade the teachers on the use of them. To exemplify, the director of C added that even the adoption of the ELP within the institution C lasted for a year to be internalized by the teachers. Correlatively, for the enhancement of on-going testing and assessment practices within the institution C and across the country, its director recommended that language testing and assessment was to be linked to a more standardized system. In addition, its director suggested that skills-based teaching was to be highlighted more, and put into use.

The viewpoints of the director of ÖZ-KUR-DER. Each of the directors was initially asked to give some information on the testing and assessment practices conducted within their institutions. Besides, the director of ÖZ-KUR-DER was also taken to semi-structured interview session in order to frame the outline better as ÖZ-KUR-DER was the Association of Private Educational Institutions and Study Centers in Turkey. In this context, its director stated that some European guidelines were to be adopted in testing and assessment procedures. To exemplify, the use of the CEFR and ELP was to be extended to a variety of educational contexts, and to be generalized across the country.

With special reference to the examination system conducted in order to present the test takers' certificates if they happened to be successful at the end of

the classes, the director of ÖZ-KUR-DER introduced that the private institution in which s/he enrolled submit a petition to the MoNE in order to organize the testing procedure after preparing the list of the test takers. If approved, the certificate examinations were held on a bimonthly basis by the MoNE. The exams were prepared by the English language teachers working for MoNE in harmony with the current curriculum. Herein, it was emphasized by the director that the teachers were not adequately qualified to prepare such kind of examinations. Besides, the weekends were chosen for the examination days. On Saturdays, a multiple-choice test was submitted to the test takers. On the other hand, an oral proficiency examination was done on Sundays. The director reported that there were 47 exam centers across the country.

Moreover, its director emphasized that the ratio of participation of the test takers was rather low as there was scarcely any candidate who went in for the examinations. The director of ÖZ-KUR-DER also added that the main reason underneath such a low level of participation might be the problematic side of the current certificate system. Herein, its director stated that the invalidity of the certificates for any other educational contexts made the situation harder. As the certificates had no validity for any further use, the test takers might prefer not to have a sit at certificate examinations.

Furthermore, it was underscored by its director that there was no standardized testing and assessment procedures followed, which might lead to compromise on quality in practice. Relatively, its director added that the quality of teaching materials and the number of teachers was adequate; however, some institutions themselves were not disposed to enhance the quality of testing and assessment practices. Due to the fact that the certificates given at the end of the English language education by all private institutions were not standardized, the attendees were also not eager to complete their classes thoroughly.

Likewise, the director of ÖZ-KUR-DER asserted that the private institutions were running to the purpose as the applications were rather goal-oriented, albeit not fulfilling the needs of the students. Hence, the results did not leave any mark as they were not meaningful in the eyes of the test takers. Besides, the English language teachers were taken to in-service training biyearly. Herein, the director recommended that this type of in-service training activities was to be more than

'holiday memories' as the teacher who joined were generally of the opinion that they had the opportunity to take a rest during those days. On the other hand, as the participants of in-service training activities were selected on a volunteer basis and with no financial support, there was a reluctance and loss of motivation, though. The director, at that point, also suggested that in-service training activities were to be conducted in cooperation with the Association, MoNE, academics and other non-governmental organizations by providing financial support for the teachers who were genuinely interested; thus, the participants could be selected more reasonably and up to the mark.

One more to note, the director argued that distance learning was adopted by the private institutions as a part of life-long learning. However, not all of the private institutions had a system of substructure running efficiently. Besides, many of them complained about the lack of teaching materials and computers. At the very same, some of the private institutions which were using the system reported that students seemed to be online although they were not sitting in front of the computers; in fact, their friends or family members made the connection in students' own usernames, and left the computers open until the end of the course hour. Thus, the students would not be marked as absent by the teachers.

Last but not least, the director of ÖZ-KUR-DER stated that the ratio of auditing was rather low as to that of formal educational settings. When the auditor or an inspector arrived, the physical structure of the private institutions and educational process were supervised, or merely asked if everything was okay when s/he came for a regular visit. Likewise, the teachers were asked to fill in the questionnaire rendered by the Ministry of National Education Data Processing Systems (MEBBIS) on an annual basis. However, the results were not sent back to the centers, or even were not announced to the teachers who were actually the participants of those questionnaires. Herein, the director recommended that the results were to be sent to the centers so as to create the reports, and present the results to the teachers as a feedback of the on-going implementations.

An overview of the directors' opinions on the utilization of the European guidelines in testing and assessment practices by selected private institutions. The viewpoints of the directors from the selected private institutions on the utilization of the European guidelines in testing and assessment practices

were highlighted by means of semi-structured interview sessions conducted face-to-face. In doing this, 6 open-ended questions were addressed in order to get more detailed information on the gist of the testing and assessment practices conducted within the institutions and their alignment to the pre-defined European guidelines.

The information was gathered on the alignment of the testing and assessment practices with the European guidelines, the instruments in use and criteria set for testing and assessment practices, the difficulties and problems mostly encountered in conducting testing and assessment practices, recommendation to enhance the quality of on-going testing and assessment practices of the private institutions together with that of the country. Accordingly, as a result of the constant-comparison analysis of the semi-structured forms, the major concepts as the needs yielded are:

1. The development of a more practical curriculum,
2. The adoption of the Framework as a basis,
3. The use of the ELP as a tool for self-assessment,
4. The validation process for language certificate examinations,
5. More qualified language teachers,
6. More effective use of the distance learning system,
7. A real auditing system for the enhancement of on-going implementations,
8. Cooperation among universities, MoNE, private institutions, ÖZ-KUR-DER and other non-governmental organizations,
9. An increase in the number of in-service teacher training activities, and
10. A standardization process in language teaching and assessment.

Taking these into account, the last chapter is composed of the conclusions drawn and the issues of discussion touched upon. Molded with an overview of the study, pedagogical implications and suggestions for further study are highlighted in detail.

Chapter 6

Conclusion and Discussion

An Overview of the Study

This dissertation scrutinized the CEFR oriented language testing and assessment practices in non-formal English language schools in Turkey. In doing this, the CEFR oriented language testing and assessment practices of 3 private institutions were probed together with the viewpoints of the directors from the selected private institutions and ÖZ-KUR-DER. Besides, the differences amidst the selected private institution in terms of implementing some European guidelines defined by the EALTA, ALTE, ILTA and AEA- Europe were investigated. Last but not least, the viewpoints of the directors of those private institutions and that of ÖZ-KUR-DER were explored in order to define the zeitgeist of the utilization of the European guidelines in testing and assessment practices of those non-formal English language schools.

In this study, the quantitative data were collected through a questionnaire on the European guidelines for establishing quality profiles in language examinations from the English language teachers at the selected private institutions, and through semi-structured interview sessions with the directors of the private institutions together with that of ÖZ-KUR-DER. The questionnaire was employed to get information on the utilization of above-mentioned European guidelines in language testing and assessment practices. Additionally, the semi-structured interview sessions were conducted in order to get information on the viewpoints of the directors of both private institutions and ÖZ-KUR-DER.

The research questions answered by this study are as follows:

1. Do the testing and assessment practices of non-formal English language schools in Turkey comply with the criteria designated by the EALTA?
2. Do the testing and assessment practices of non-formal English language schools in Turkey correspond to the standards set by the ALTE?
3. Do the testing and assessment practices of non-formal English language schools in Turkey fit the guidelines assigned by ILTA?

4. What is the role of testing and assessment in Turkey's system of education in the light of the standards set by the AEA-Europe?

5. What is the general paradigm of a sample of leading professionals from selected non-formal English language schools in Turkey (i.e. decision-makers, testing office, English language teachers) on the implementation of testing and assessment procedures as defined by the European guidelines?

a. Do the testing and assessment practices of selected non-formal English language schools in Turkey differ from each other within the scope of pre-determined European guidelines?

b. What are the viewpoints of the directors from the selected private institutions and ÖZ-KUR-DER on the utilization of the European guidelines in testing and assessment practices?

Besides, the viewpoints of the directors from the selected private institutions on the utilization of the European guidelines in testing and assessment practices were highlighted by means of semi-structured interview sessions conducted face-to-face. Herein, 6 questions were addressed by the researcher to get qualitative data on the gist of the testing and assessment practices conducted within the institutions and their alignment to the pre-defined European guidelines. Accordingly, these questions were listed as:

1. Please provide some information on the testing and assessment practices conducted within your institution(s).

2. Are these practices aligned with any European standards? If yes, please provide some information about those standards.

3. Please provide some information about the instruments and the criteria set for testing and assessment practices.

4. Please provide some information on the difficulties and problems mostly encountered in conducting testing and assessment practices.

5. Please provide some recommendations in order to enhance the on-going testing and assessment practices within your institution(s).

6. Please provide some recommendations in order to enhance the on-going testing and assessment practices across the country.

An in-depth analysis of the results was applied in a three-way alternate: (1) data collected from a 5-point-Likert type scale, and demographic information gathered from the English language teachers (40 in total, 28 female and 12 male, working at the selected private institutions also as the test-item developers); (2) data gathered from the semi-structured interview sessions conducted with the directors of the selected private institutions (3 directors in sum); (3) data gathered from the semi-structured interview session conducted with the director of ÖZ-KUR-DER.

In this section, the findings of the study are reported and discussed. In the sequel, pedagogical implications of the study are presented. Besides, some suggestions for further research are given. Lastly, the summary of the results is supplied with a conclusion part.

Discussion of the Results

The aim of this dissertation, as has been stated, is to arrive at an understanding of the on-going testing and assessment practices of the English language schools in Turkey, which are serving as private non-formal educational institutions, in terms of some European standards. These European standards are grounded upon the EALTA, ALTE, ILTA and AEA- Europe, taking the Framework as the core element. In this section, the results obtained from data analysis are reviewed and discussed in three sub-parts:

1. The utilization of the European guidelines in testing and assessment practices by selected private institutions,
2. The viewpoints of the directors from the selected private institutions on the utilization of the European guidelines in testing and assessment practices,
3. The viewpoints of the director of ÖZ-KUR-DER on the utilization of the European guidelines in testing and assessment practices.

The utilization of the European guidelines in testing and assessment practices by selected private institutions. The data regarding the utilization of the European guidelines in testing and assessment practices by selected private institutions have yielded that even the most prominent English language schools which are renowned for quality in learning English in Turkey with the highest course attendee capacity and with the highest number of branches across the country have

not embraced these European guidelines in language testing and assessment thoroughly. Even more, they assert that they have adopted the Framework as the fundamental basis for language testing and assessment procedures conducted, albeit inefficiently. Besides, it is reported by the findings of this study that the English language teachers, who are also test (-item) developers at those private institutions, are well aware of the importance of the Framework, and the significance of the alignment of the tests in use to the Framework. In this context, Coste (2007) signifies the notable influence of the Framework on language testing and assessment; thus, the alignment of the language tests to the Framework has drawn more attention than the others aspects.

Correlatively, within the boundaries of the considerations for test development in national or institutional testing units or centers, the EALTA (2006) seeks for answers to seven key questions, one of which is grounded upon the process of 'linkage to the CEFR'. However, with a scarcity of empirical studies regarding the utilization of the EALTA Guidelines in language testing and assessment (Alderson & Banerjee, 2008; Alderson, 2010; Erickson & Figueras, 2010; De Jong & Zheng, 2011; Kavakli & Arslan, 2017; Toncheva, Zlateva & John, 2017), it is noted that the EALTA Guidelines have not been solely applied for the language testing and assessment practices of non-formal educational settings before. However, as noted by Alderson (2010), the EALTA Guidelines are assumed to be used to 'frame a validity study' (p. 63). Taking this as the starting point, it is reported by the findings of this study that the English language teachers who are also test (-item) developers at those private institutions are not well informed whether there is a publicly available report on the alignment of the tests in use to the Framework (M= 3.75). Correlatively, not all of the tests in use correspond to the procedures recommended in the Framework by the Manuals and Reference Supplement (M= 3.90). It is also reported that the ELP is to some extent in use as a self-assessment tool in these selected private institutions (M= 3.98) though the ELP is implemented widely around the world via its free access in many languages by learners and teachers (CoE, 2011; Little, 2005; Mirici, 2008; Schaerer, 2005). At the very same, the ELP is directly linked to the Framework, providing a common basis for evidencing language syllabi, curriculum guidelines, textbooks, language examinations and the like beyond Europe (CoE, 1998; 2001).

Moreover, the EALTA also seeks for answers to the questions entailing test purpose and specification, test design and item writing, quality control and test analyses, test administration, review and washback. Accordingly, the concerned stakeholders such as learners, teachers, test (-item) developers, directors and general public are made aware of the clarifications in testing and assessment practices. Herein, Alderson (2007) suggests that the EALTA is the organization which is acting as an equivalent mechanism to validate the claims of the examination providers. However, the overall results have showed that the lowest mean score was estimated upon the utilization of the EALTA Guidelines in language testing and assessment practices (M= 3.73).

Additionally, the ALTE defines the characteristics of the examinations by means of the Code of Practice and Minimum Standards through test construction, administration and logistics, marking and grading, test analysis, item writing, test production and communication with stakeholders. In this study, the data gathered from the English language teachers nestle each of the components aforementioned. Although the ALTE Code of Practice is proposed as a cadre for monitoring professional standards in language testing and assessment (Saville, 2005), the second lowest mean score is estimated upon the utilization of the ALTE Code of Practice in language testing and assessment practices (M= 3.74). To elaborate each of the components, the mean score estimated for the sub-component of administration and logistics falls behind of all (M= 3.53). It is followed by that of communication with the stakeholders (M= 3.78), test analysis (M= 3.82), item writing (M= 3.86), test production (M= 3.88), test construction (M= 3.89), and marking and grading (M= 3.89) respectively. The ALTE Code of Practice has been proposed as one of the European criteria to enable test fairness and to set priorities (Xi, 2010); however, the selected private institutions lag behind in implementing the ALTE Code of Practice effectively in language testing and assessment practices. By the same token, the ALTE is the association that has introduced the Manual for Language Test Development and Examining as a complementary document for the 'Manual for Relating Language Examinations to the CEFR' (CoE, 2009a) on behalf of the Language Policy Division of the CoE to be used with the CEFR effectively within their own contexts, and by their own objectives (CoE, 2011).

To note more, the AEA- Europe is elaborated as a platform in which developments of educational assessment within Europe are discussed. Besides, the 'European Framework of Standards for Educational Assessment' (AEA- Europe, 2012) developed by the AEA- Europe is highlighted. The data gathered from the English language teachers in the adoption of the Framework of the AEA- Europe in language testing and assessment practices are probed in two sub-components, namely the guiding principles and instrument. By the results of the data analysis, it is reported that the guiding principles are applied more ($M= 3.86$) than the instrument ($M= 3.76$) although overall estimates regarding the adoption of the AEA- Europe's Framework is cumulatively low ($M= 3.81$). Herein, it is to be noted that the lowest mean score of the guiding principles is estimated for item no. 65 ($M= 3.45$), which is about the goodness of the test takers as the individuals who are taking the tests if aligned with the United Nations Convention on the Rights of the Child (UN, 1990). Accordingly, individual's place in the assessment procedure is expected to be guaranteed by the declaration of the UN, confirming that everyone is entitled to all rights asserted without any distinction of any kind, such as race, ethnicity, language, gender, or any other status. However, it is stipulated by the findings of this study that a big majority of the English language teachers as the participants of the study are not well-aware of what it is actually about as they have noted themselves as 'not sure' in reply to the aforementioned test item ($P= 42.5\%$).

On the other hand, it is noted by the findings regarding the instrument by AEA- Europe that the lowest mean score is estimated on the use of standardized tests within selected private institutions ($M= 3.45$). In effect, a Reference Supplement to the Manual for Relating Examinations to the CEFR has been introduced (Banerjee, 2004; Eckes, 2009; Kaftandijeva, 2004; Verhelst; 2004a, -b, -c, -d) to enable standardization in developing tests, and aligning them to the Framework. To note more, it is reported by the findings of this study that summative assessment is the type of assessment which is most generally applied in the selected private institutions ($M= 4.00$). It is followed by the implementations of formative assessment ($M= 3.82$), and those of performance assessment ($M= 3.73$). Herein, Spinelli (2007) has suggested informal assessment as an authentic solution to the need for formative assessment in order to involve individual's learning styles and personal challenges into the process; thus, teachers can track the on-going

educational process more regularly, and often by taking students' snapshots throughout the process.

Besides, ILTA offers a number of basic tenets for its members by identifying the responsibilities of test designers, test item writers, institutions involved, stakeholders as the test result users, and test takers. Herein, ILTA buoys ethical standards by means of the 'Code of Ethics' (ILTA, 2000), and principles to enable good testing practice in all situations thanks to the 'Guidelines for Practice' (ILTA, 2007). Accordingly, the ILTA Guidelines for Practice, which are taken into the questionnaire as the standards for establishing quality profiles in examinations, are embedded in two sub-parts, namely responsibilities of the test designers and test writers, and responsibilities of the test takers. The highest mean score of all European standards above-mentioned in detail is estimated on the adoption of the ILTA Guidelines for Practice in language testing and assessment practices ($M=4.03$). To elaborate, the first component has the mean score of 4.05 whereas the latter has that of 3.99 out of 5. Therefore, it is stipulated that test designers are well aware of their own responsibilities together with those of test takers as the English language teachers as the test (-item) developers and/or designers at selected private institutions are assumed to inform the candidates of their rights. It might be due to the changing nature of the course characteristics, and the features of the instructors and learners (Brown & Bailey, 2008).

To briefly sum up the utilization of the European guidelines in testing and assessment practices by selected private institutions, the private institution B has outscored the others in each of the European guidelines. This might be due to the fact that the private institution B is reported by its director to adopt the Framework, and use the ELP as a tool in classroom-based assessment; thus, it is more acquainted with the operational procedures of the CEFR.

The viewpoints of the directors from the selected private institutions on the utilization of the European guidelines in testing and assessment practices.

The general paradigm of a sample of leading professionals from a number of non-formal English language schools in Turkey on the implementation of testing and assessment procedures as defined by the European guidelines is drawn taking the views of the decision-makers and English language teachers, who are also working as test (-item) developers at the same institutions. Herein, the results are presented

in a two-way alternate. Firstly, the overall estimations regarding the exploitation of all European standards by selected private institutions are reported at one hand. Secondly, the viewpoints of the directors from the selected private institutions are addressed by their own answers gathered from the semi-structured interview sessions to the accompaniment of 6 questions. Additionally, the viewpoints of the director of ÖZ-KUR-DER are also noted in the following section in order to frame the outline better, and to enable a triangulation by the answers gathered from the English language teachers, the directors of the selected private institutions and the director of ÖZ-KUR-DER.

Accordingly, the data regarding the utilization of the European guidelines in testing and assessment practices gathered from the directors of those selected private institutions have yielded that similar types of assessment formats are in use, which are mostly summative. Test takers are provided with contemporary self-assessment tools, such as the ELP to some extent. For the private institution B, it is the classroom-based assessment tool. However, for the private institution A and C, there are some restrictions in use, such as age and language proficiency level. However, the ELP is the fundamental tool for learners to keep record their own learning by themselves (Sarıçoban, 2011); therefore, the recognition and implementation of the ELP is a necessity of the time, albeit not a choice. Besides, the placement tests are conducted after pre-defined hours of lecture to define the language proficiency levels of the learners. Herein, the RLDs of the Framework is applied as the proficiency levels are named accordingly. The RLDs of the Framework are actually more meaningful than just labelling the numerical data. Herein, Trim (2005) states that the descriptors of the Framework have been of interest of the educational authorities, learners, even employers who are to some extent in need of situating their language qualifications. Therefore, the Framework has been approved as being meaningful in the domains of testing, assessment and certification (CoE, 2005b). However, there mushrooms a risk if the Framework is used for assessment without any calibration. Yet already, North (2014) asserts that this is the secret underneath the CEFR as there are variety of interpretations of it. To note more, the descriptors have undergone a recent revision by the CEFR Companion Volume, through which Pre- A1 level of language proficiency is newly added to the already existing RLDs (CoE, 2017).

As these private institutions are rendering English language education under the frame of NFE, they are mostly regarded as trading houses which are merchandizing education. Therefore, the student(s) enrolled in such kind of private institution(s) are well-aware of the fact that it is the identity of the institution(s) which is protected, albeit not that of student(s). Reviewing the recognition of the policies and practices in NFE of the EU, Bjornavold (2000) suggests that contextual nature of learning, identification of methodological requirements for assessing non-formal learning, and institutional requirements together with the political stance are to be reconsidered in conducting educational activities on a non-formal basis. Similarly, Eaton (2010) has made a differentiation amidst formal, non-formal and informal learning settings of Canada regarding literacy, essential skills and language learning in the light of the Framework due to the fact that each types of learning encompasses different types of requirements within.

Besides, some problematic issues blossom as there are no standards in language testing and assessment practices of the selected private institutions. Therefore, someone with a proficiency level of B1 might be regarded as proficient at the level of A2 by another institution. According to the views of the directors from selected private institutions, it is reported that current testing and assessment practices are to be linked to a more standardized system. So, the problem is setting standards for quality. However, it is to be noted that setting standards is not the same with adopting standardization due to the fact that standardization refers to settings things in completely the same way (Sleeter & Carmona, 2017). Even so, such kind of standardized tests should at least be laced with some alternative assessment measures (Menken, 2008). Herein, Jones (2009) asserts that standard-setting could be enabled through the use of the Framework.

The viewpoints of the director of ÖZ-KUR-DER on the utilization of the European guidelines in testing and assessment practices. The viewpoints of the director of ÖZ-KUR-DER are also noted in order to frame the outline better, and to enable a triangulation by the answers gathered from the English language teachers, the directors of the selected private institutions and the director of ÖZ-KUR-DER. Accordingly, the data regarding the utilization of the European guidelines in testing and assessment practices gathered from the director of ÖZ-KUR-DER have yielded that the ratio of participation of the test takers is noted as rather low as

there is scarcely any candidate who goes in for the examinations. Herein, this low level of participation is attributed to the invalidity of the current certificates for any further educational use. However, the examinations are assumed to be conducted in order to certify the successful ones formally, as a sign of completion of a program which is either formal, or non-formal (Scheerens, Glas & Thomas, 2003). As the results gathered by the language certificate examinations are not fully applicable, the certificates are regarded as a piece of papers. At that point, setting standards in language testing and assessment practices blossoms as a need according to the viewpoints of the director of ÖZ-KUR-DER.

Correlatively, it is pointed out by the director of ÖZ-KUR-DER that the quality of testing and assessment practices is enhanced through developing teacher qualifications due to the adoption of more appropriate testing and assessment activities. It, then, yields to teachers' much better understanding of the process together with the learners' much better internalization of the procedure (Lambert & Lines, 2000). In a similar vein, the director of ÖZ-KUR-DER also states that these types of private institutions should not be regarded as free-of-charge certificate deliverers. Since if it is the case, the learners most probably focus on the end-of-course examinations more than the process. However, focusing on examinations is not preferred, instead assessment for learning is expected in order to forge a link between the learners and institutions (San, 2016). Besides, exam-oriented study of the learners may cause teachers not to apply contemporary approaches in a classroom environment; thus, the classroom applications become rather goal-oriented, albeit not fulfilling the needs of the students.

One more note, the director of ÖZ-KUR-DER emphasizes that the ratio of auditing is rather low as to that of formal educational settings. Even more, when the auditor or an inspector arrives, the physical structure of the private institutions and educational process are checked superficially. Likewise, the teachers are asked to fill in the questionnaire rendered by MEBBIS on an annual basis. However, the results are not sent back to the centers, or even are not announced to the teachers. Herein, the director underlines the significance of these results, regarding them as the feedback of the on-going implementations due to the fact that the assessment essentials not only covers developing learning goals, objectives and planning for

assessment, but it also embodies assessing the assessment program (Palomba & Banta, 1999).

Pedagogical Implications

The results of this study suggest that the testing and assessment practices of the selected private institutions rendering non-formal English language education have not framed by the CEFR thoroughly. The majority of the English language teachers assumes that the testing and assessment practices conducted within their institutions do not fully cover the principles set by the guidelines of the EALTA, ALTE, ILTA and AEA- Europe. Besides, the tenets of the CEFR which have been adopted by these private institutions are not put into practice effectively although the directors emphasize the significance of a more practical curriculum enabled through the adoption of the Framework. Correlatively, the ELP blossoms as a tool for self-assessment applied in classroom-based language assessment. However, the results of this study yield that there is no authorized standard set for the use of the ELP in foreign language classrooms, or the standards are set by the private institutions themselves (e.g. learners with 15+ age, or learners with A2+ level of language proficiency). However, utilizing European guidelines requires establishing quality standards in language testing and assessment practices (Kavaklı, 2017b).

Moreover, the results of this study reveal that one of the major problems encountered in the testing and assessment practices of non-formal educational settings is the invalidity of the certificates given at the end of the foreign language education if marked as successful by language certificate examination. Therefore, a validation process for language examinations mushrooms as a need for such settings, which could be enabled by means of a cooperation among the allies: universities, MoNE, private institutions, ÖZ-KUR-DER, public education centers, governmental agencies and other non-governmental organizations. Otherwise, the learners are to take additional courses to take another certificate defining their language proficiency skills by means of an internationally recognized and accredited language test.

In this context, it is also reported that the language certificate examinations are prepared by the language teachers themselves. Therefore, there springs the need for more qualified language teachers, and that of an increase in the number of

in-service teacher training activities in tow. The need for more qualified teachers can be satisfied with the adoption of a revised ELT curriculum based on the Framework. Since, as the current reality of the ELT professionals, the Framework is now more than just being 'common' and 'European', albeit internationally recognized worldwide (Mirici & Kavaklı, 2017). It is also recommended by the director of ÖZ-KUR-DER that in-service teacher training activities should be given utmost importance to make the teachers internalize the Framework better through the instrument of a co-operation between the MoNE, and Association of Private Educational Institutions and Study Centers in Turkey. Herein, the fifth principle of the ILTA Code of Ethics (2000) embarks on the enhancement of the language testers' professional knowledge, and sharing it with other language professionals through co-operation by keeping themselves up-to-date with the latest developments and novelties in the field, and applying them for the goodness of their test takers.

With regard to the problems encountered regarding the effective use of the distance learning system as a part of lifelong learning in non-formal educational settings, it is reported that being online at the pre-defined course hours is enough to go in for the language certificate examination, paving the way towards an unfair competition between the test takers. Besides, no standards have been previously set for the functioning of the system efficiently. Even more, some non-formal educational settings are lacking in related materials and background systems required. Hence, the understanding of 'computer-assisted language learning' (Levy, 1997) should be enhanced to be more precise. Besides, the ratio of auditing of non-formal educational settings is rather low as to that of formal educational environments. It is supported by the results of this study that there is a need for a real auditing of the on-going testing and assessment practices of non-formal educational settings. Therefore, it is essential to provide feedback for the centers together with the teachers with regard to the assessing of the assessment program.

In the light of these, it mushrooms as a crystal-clear fact that there is a need for a procedure of setting standards in language testing and assessment practices. Herein, for setting standards in order to enhance the quality of the on-going testing and assessment practices, the Framework can be applied as a guide (Jones, 2009). In order to conduct such kind of standardized examinations, it is suggested by the

ILTA Guidelines for Practice (2007) that the test preparation stage should be firstly depended upon the language testing theory which is currently in use, and for those who are non-native speakers of the language being tested, someone with a high level of proficiency in the aforementioned language is to be employed to check the items written. For standardized assessments at high-stakes, the test difficulty and score comparability is to be investigated in order to ensure fairness among test takers, and the results are to be interpreted and reported accurately (ILTA, 2007). Besides, all test takers should be provided with satisfactory information about the procedure. The results gained at the end should be announced correctly and put in the data-base after a continuous quality control analysis. It is also to be noted that if there is more than one form, inter-form reliability is to be calculated and published, as well.

In this context, including standardized tests into the procedure should not be regarded as the constant use of the same types of assessment all the time. Since standardized tests should also be embellished with some alternative assessment measures, as well (Menken, 2008). In the same vein, it is recommended that formative assessment should enhance learning by providing feedback for both teachers and learners together with the opportunity for self-evaluation (Walvoord & Anderson, 2010). Additionally, skills-based assessment should be taken to the core as the Framework, itself, defines each of the language skills in a comprehensive way by means of 'can-do' statements. Halbherr, Schlienger and Piendl (2014) emphasize that assessment practices should be molded in reply to globalization around the world; therefore, assessment for a digital world is to be revised and re-arranged in accordance with the Framework. Additionally, the assessment types which support learning, and fit for a European environment are expected to be addressed, gathering the on-going traditions in assessment and new forms of approaches together (AEA- Europe, 2012). Correlatively, "explicitly or implicitly defined in opposition to traditional externally set and assessed large scale formal examinations" (Davison & Leung, 2009, p. 395), LOA is suggested grounded upon a socio-cognitive model of language learning propounded by the Framework (Jones & Saville, 2014) to signal the concept of fitness-for-purpose.

To sum up, a more centralized certificate examination should be applied in all of the non-formal English language learning and teaching settings. Thus,

common testing and assessment practices should be conducted for all types of non-formal educational settings, such as public education centers, private institutions, training courses catered by the municipalities, and other governmental agencies. Herein, the Framework is to be taken as the baseline for enhancing quality standards in language testing and assessment practices. In addition, the certificates should be valid for further use. Therefore, the validation of those certificates may be enabled through adding some extra points for those who are going in for large-scale examinations, or passing English language proficiency examinations sit for preparatory class exemption without taking it. Embracing a vast majority of ELLs and contributing to a large part of the Turkish education economy, non-formal educational settings are of utmost importance. It is reported by the Directorate General for Private Education Institutions of MoNE that 67.000 students are learning English through out-of-school education, meaning that providing an amount of approximately 1500 Turkish Liras per person, 100.000.000 Turkish Liras are spent each year to learn English at private institutions (Karaboğa, 2013). Therefore, as indicated by the results of this study, dissemination of the European guidelines is to be encouraged in the language testing and assessment practices of those private institutions. Herein to emphasize, long-term meaningful effects are to be reckoned until acceptable results are achieved in order to ensure 'no tissue rejection' (Holliday, 1992).

Suggestions for Further Studies

This study aims to arrive at an understanding of the on-going testing and assessment practices of the English language schools in Turkey, which are serving as private non-formal educational institutions, in terms of some European standards. Accordingly, the utilization of the European guidelines in testing and assessment practices by selected private institutions is embarked on together with the viewpoints gathered from the directors of the selected private institutions, and with the viewpoints gathered from the director of ÖZ-KUR-DER.

In this sense, 3 private institutions rendering non-formal English language education are included in this study although they are the most commonly preferred, widely-known and influential language learning centers of private education sector in Turkey. For further research, this number could be increased encompassing different types of non-formal English language learning centers, such as public

education centers, other private institutions, training courses catered by the municipalities, and other types of governmental agencies. To add more, the test formats in use by the selected private institutions could be analyzed with regard to the essentials of testing and assessment, such as validity, reliability, etc.

Additionally, the students are not included within this study, assuming that it would be hard to control such a wide range of variables all at once. Further research could nestle the learners as a separate variable by taking their viewpoints for the enhancement of the on-going testing and assessment practices. Besides, the focus of attention in this study is the Turkish context, where English is taught as a foreign language. Herein, targeting at some other European countries that have adopted the Framework could expand the circle. As they are more familiar with the content of the CEFR, such a comparative study could probably lead researchers to reach manifold conclusions. In any case, examining a wider range of curricula under the influence of the CEFR, namely not school-based and non-formal educational settings, would surely broaden the viewpoints as there is a scarcity of empirical studies in this scope.

Conclusion

This mixed-methods research, laced with both qualitative and quantitative data, aimed at arriving at an understanding of the on-going testing and assessment practices of three institutionalized private English language schools offering non-formal English language education in their branches in all of the major cities in Turkey. Grounded upon the European standards for language testing and assessment, the qualitative data were gathered from the directors of these private institutions and the director of ÖZ-KUR-DER within the scope of general information about the institution, the running of the on-going testing and assessment practices laced with numeric data on the number of teachers, test (-item) developers and students, and the difficulties and problems encountered in the implementation of the testing and assessment practices together with the recommendations for further improvement whereas the quantitative data were gathered from teachers who were also working as test (-item) developers at those private institutions regarding the utilization of the European guidelines in testing and assessment practices within the private institutions they were working at.

Accordingly, the findings revealed that the language testing and assessment practices of those selected private institutions were found to be appropriate only to some extent. The English language teachers in this study needed to develop their language testing and assessment skills in relation with the Framework. Amidst the European guidelines of language testing and assessment set by the EALTA, ALTE, ILTA and AEA- Europe, the EALTA Guidelines were marked as applied least regarding their utilization by the English language teachers, who were also working as test (-item) developers at the selected private institutions. This might be attributed to the robust nature of the procedures of the quality control and test analyses, review and washback, test design and item writing, and linkage to the CEFR. Herein, it was noted that the procedure of review and washback was not attentively considered, which might indicate that the significance of reporting the expected test effects either positive or negative was not imprinted well in English language teachers' minds. However, providing feedback is essential for both decision-makers and program reviewers in order to enhance the quality of educational assessment, and to evaluate programs. When it comes to the phases of test construction and production ascertained by the ALTE, the findings showed that the procedure of test analysis was fit for purpose, yet not covered smoothly. Thus, it paved the way towards the emergence of some application-oriented problems regarding administration and logistics. Henceforth, it might be concluded that the English language teachers, who were also working as test (-item) developers at the selected private institutions were not provided with sufficient guidance and support during the testing and assessment processes.

The findings of this study also revealed that the responsibilities of the test designers and/or writers were given much importance than those of test takers with regard to the exploitation of the ILTA Guidelines for Practice. Therefore, it could be stipulated that the test designers and/or writers well understood and clearly applied their responsibilities during the testing and assessment procedures. Adopting themselves as the main beneficiaries, the English language teachers have not given due weight in order to guarantee test takers' rights since the assessment process is inscribed more to the test administrators and developers more than test takers. With special concern to the AEA- Europe's Framework, the findings yielded that the guiding principles were applied merely to some extent by the English language

teachers, who were also working as test (-item) developers at the selected private institutions. Herein, the English language teachers admitted that they suffered from using traditional assessment techniques more than innovative ones although the Framework, itself, did focus on educational assessment supporting learning. This might indicate that new forms of assessment which fit for a European environment are not adequately placed emphasis. It also seems that disseminating quality in educational assessment for the development of quality in educational assessment with a European perspective have blossomed as a need for all of the private institutions rendering English language education in a non-formal way. In this context, the English language teachers might have experienced role models or mentors in order to grasp the gist of the Framework, and to use it in their classes. Or, they might take the advantage of further in-service teacher training facilities; however, the findings revealed that in-service teacher education fell short of providing teachers with an awareness of the CEFR and ELP, and therefore, leading them to continue with the habit of on-going reiteration of the same old story without the reconceptualization of the current EFL curriculum in use.

Concerning the fundamental assessment principles, the findings yielded that a certain degree of credibility was not appropriately reflected by the results gained through the tests in use. Therefore, the certificates given at the end of the courses were not marked as valid due to the fact that the quality aspects as the cornerstones of a professional assessment were not covered thoroughly. In this context, the cooperation amidst the allies labelled as the MoNE, universities, private institutions and/or courses, ÖZ-KUR-DER and other non-governmental organizations is assumed to contribute to the development and use of standardized language tests to ensure the validity of language certificates rendered. Besides, standard requirements, methods and samples of evidence as the sub-components of the instrument set by the AEA- Europe were stipulated not to be sufficiently addressed by means of observations and verifications. This situation reveals that the design of the assessment procedure does not properly represent the content which is covered by knowledge, skills and other attributes, and the setting in which the assessment is going to take place. For the evaluation and next iteration phase, the results are expected to embrace their further use for other educational cases; however, it was yielded by the findings of this study that the concept of next iteration was not fully

understood either to develop a new form of assessment, or to improve the already existing one within the scope of the European standards touched upon above. Herein, a more robust auditing system is needed in order to enhance the quality of language testing and assessment practices in non-formal educational settings.

Another interesting finding was that the proper interpretations of the test results were not made; thus, the rationale behind the test outcomes was not utterly comprehended by the test takers. This might be attributed to the indetermination of the harmony between the testing criteria and test characteristics by defining the goal of assessment, and entering the construct and function into the testing and assessment process. Correlatively, the English language teachers, who were also working as test (-item) developers at the selected private institutions did not have a good grasp of the concept of communication with the stakeholders as they recapped this process with barely announcing the test results. This might be attributed to the inappropriate use of the information sharing process for the interpretation of the results more prominently. However, this communication process does not only embrace the announcement of the results but also includes sharing information on the context, purpose and use of the examination.

To sum up, in this study, the current language testing and assessment practices in non-formal educational settings, as the arteries of Turkish education economy, has been discussed to improve the quality by the exploitation of the CEFR. Accordingly, it could be stipulated that even the most prominent English language schools in Turkey do not apply European guidelines in language testing and assessment thoroughly. Moreover, it mushrooms by the viewpoints of the directors of these private institutions and ÖZ-KUR-DER that the Framework is not adequately covered in testing and assessment practices conducted by the English language schools serving as non-formal educational settings. Correlatively, it is reported by the aforementioned directors that as the results gathered by language certificate examinations are not fully applicable and valid for further educational use, the number of test takers gets much lower; henceforth, setting standards in language testing and assessment practices blossoms as a need.

In conclusion, formal educational settings have generally been at the core of studies conducted in the field of language testing and assessment. The CEFR oriented testing and assessment practices have also been wheeled around formal

educational settings. Contrary to the ordinary, the testing and assessment practices of non-formal private institutions are taken as the core instructional context within this study. Bridging the gap in the literature, this study opens up a new understanding of the utilization of some European standards in language testing and assessment practices by selected private institutions rendering English language education. This study also contributes to foreign language research in the field of testing and assessment, highlighting the role of setting standards in testing and assessment as a need for subsequent practices. Besides, this study underscores the significance of a valid certification system for defining foreign language proficiency; thus, the test takers can make a better sense of the learning process. Last but not least, the results of this study are expected to lend assistance to different types of audiences: English language teachers, test (-item) developers, the directors of the private institutions, public enterprises and the directors of other non-governmental organizations.

References

- Akin, G. (2016). Evaluation of national foreign language test in Turkey. *Asian Journal of Educational Research*, 4(3), 11-21.
- Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal (MLJ)*, 91(4), 659–663.
- Alderson, J. C. (2010). A survey of aviation English tests. *Language Testing*, 27(1), 51-72.
- Alderson, J.C., & Banerjee, J. (2008). *EALTA's guidelines for good practice: A test of implementation*. Paper presented at the 5th Annual Conference of the European Association for Language Testing and Assessment. Athens, Greece, 8- 11 May, 2008. [On-line: <http://www.ealta.eu.org/conference/2008/programme.htm>, Retrieved on 15 February, 2017.]
- Alderson, J. C., & Huhta, A. (2005). The development of a suite of diagnostic tests based on the common European framework. *Language Testing*, 22, 301–320.
- Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analyzing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR Construct Project. *Language Assessment Quarterly*, 3(1), 3-30.
- Association for Educational Assessment in Europe (AEA- Europe). (2012). *European Framework of Standards for Educational Assessment (Version 1.0)*. Rome: Edizioni Nuova Cultura.
- Association for Educational Assessment in Europe (AEA- Europe). (2013). *Constitution for the association for educational assessment in Europe*. [On-line: <https://www.aea-europe.net/about-us/aims-and-objectives/>, Retrieved on 16 June, 2016.]
- Association of Language Testers in Europe (ALTE). (2012). *Constitution for the association of language testers in Europe*. [On-line: <http://www.alte.org/docs/constitution-2012.pdf>, Retrieved on 16 June, 2016.]
- Banerjee, J. (2004). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the CEF: Section D: Qualitative analysis methods*. Strasbourg: Language Policy Division.
- Bjornavold, J. (2000). *Making learning visible: Identification, assessment and recognition of non-formal learning in Europe*. Luxembourg: European Communities.
- Bogdan, R. C., & Biklen, S. K. (2003). *Qualitative research of education: An introductory to theories and methods* (4th edition). Boston: Allyn and Bacon.

- Bonnet, G. (2007). The CEFR and education policies in Europe. *The Modern Language Journal (MLJ)*, 91(4), 669-672.
- Brown, J. D., & Bailey, K. M. (2008). Language testing courses: What are they in 2007? *Language Testing*, 25(3), 349-383.
- Byram, M., & Parmenter, L. (2012). *The common European framework of reference: The globalization of language education policy*. Bristol: Multilingual Matters.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, 1-47.
- Candy, P. (1991). *Self-direction for lifelong learning*. San Francisco: Jossey-Bass Inc.
- Chamot, A. U., & O'Malley, M. (1987). The cognitive academic language learning approach: A Bridge to the mainstream. *TESOL Quarterly*, 21(2), 227-250.
- Chastain, K. (1988). *Developing second language skills: Theory and practice* (3rd edition). Florida: Harcourt Brace Jovanovich.
- Choi, I.-C. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing*, 25(1), 39-62.
- Colardyn, D. (Ed.). (2002). *Lifelong learning: Which ways forward?* Utrecht: Lemma.
- Coombs, P. H., Prosser, C., & Ahmed, M. (1973). *New paths to learning for rural children and youth*. New York, NY: International Council for Educational Development.
- Coste, D. (2007). *Contextualizing uses of the common European framework of reference for languages*. Paper presented at Council of Europe Policy Forum on use of the CEFR, Strasbourg, France, 6- 8 February, 2007. [On-line: <https://rm.coe.int/contextualising-uses-of-the-common-european-framework-of-reference-for/16805ab765>, Retrieved on 10 October, 2016.]
- Council of Europe (CoE). (1996). *Users' guide for examiners*. Strasbourg: Language Policy Division.
- Council of Europe (CoE). (1998). *Learner autonomy in modern languages*. Strasbourg: Editions of Council of Europe.
- Council of Europe (CoE). (2000). *Resolution on the European language portfolio*. Adopted at the 20th session of the Standing conference of the ministers of education of the Council of Europe, Cracow, Poland, 15-17 October, 2000. [On-line: <http://culture.coe.int/portfolio>, Retrieved on 7 February, 2017.]
- Council of Europe (CoE). (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

- Council of Europe (CoE). (2005a). *Reference level descriptions for national and regional languages (RLD) – Guide for the production of RLD*, (Version 2). Strasbourg: Language Policy Division.
- Council of Europe (CoE). (2005b). *Survey on the use of the common European framework of reference for languages (CEFR): Synthesis of Results*. [On-line: <https://www.coe.int/t/dg4/linguistic/source/surveyresults.pdf>, Retrieved on 20 July, 2017.]
- Council of Europe (CoE). (2006). *Plurilingual education in Europe. 50 years of international co-operation*. Strasbourg: Language Policy Division.
- Council of Europe (CoE). (2009a). *Relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment (CEFR): A manual*. Strasbourg: Language Policy Division.
- Council of Europe (CoE). (2009b). *Relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment (CEFR): Further material on maintaining standards across languages, contexts and administrations by exploiting teacher judgment and IRT scaling*. Strasbourg: Language Policy Division.
- Council of Europe (CoE). (2011). *Manual for language test development and examining: For use with the CEFR*. Strasbourg: Language Policy Division.
- Council of Europe (CoE). (2017). *Companion volume with new descriptors*. Strasbourg: Council of Europe.
- Council of Higher Education (CoHE). (2007). *Türkiye'nin yükseköğretim stratejisi* [Higher education strategy of Turkey]. Ankara: Council of Higher Education. [On-line: http://www.yok.gov.tr/documents/10279/30217/yok_strateji_kitabi/27077070-cb13-487-0-aba1-6742db37696b, Retrieved on 25 June, 2017.]
- Council of Higher Education (CoHE). (2017). *Announcement by the equivalence office*. Ankara: Council of Higher Education. [On-line: http://yok.gov.tr/web/en-denklik-birimi/home/-/asset_publisher/YPDsy2oQ2JMA/content/about-yokdil-exam-grades-28-04-2017-?redirect=http%3A%2F%2Fyok.gov.tr%2Fweb%2Fen-denklik-birimi%2Fhome%3Fp_p_id%3D101_INSTANCE_YPDsy2oQ2JMA%26p_p_lifecycle%3D0%26p_p_state%3Dnormal%26p_p_mode%3Dview%26p_p_col_id%3Dcolumn-2%26p_p_col_count%3D2, Retrieved on 22 June, 2017.]
- Creswell, J., & Plano Clark, V. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.
- Cumming, A. (2009). Language assessment in education: Tests, curricula, and teaching. *Annual Review of Applied Linguistics*, 29, 90–100.
- Cummins, P. W., & Davesne, C. (2009). Using electronic portfolios for second language assessment. *The Modern Language Journal (MLJ)*, 93(1), 848-867.

- Dancey, C. P., & Reidy, J. (2004). *Statistics without Maths for psychology: Using SPSS for windows*. London, UK: Prentice Hall.
- Davidson, F., & Fulcher, G. (2007). The common European framework of reference (CEFR) and the design of language tests: A matter of effect. *Language Teaching*, 40(3), 231-241.
- Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing*, 25(3), 327-347.
- Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly*, 43(3), 393-415.
- De Jong, J.H.A.L. (2004). *Comparing the psycholinguistic and the communicative paradigm of language proficiency*. Paper presented at the International Workshop Psycholinguistic and Psychometric Aspects of Language Assessment in the Common European Framework of Reference for Languages, University of Amsterdam, The Netherlands, 13- 14 February, 2004. [On-line: https://books.google.com.tr/books?id=DIZ-t94wBDwC&pg=PA177&lpg=PA177&dq=Comparing+the+psycholinguistic+and+the+communicative+paradigm+of+language+proficiency.&source=bl&ots=JleW6IWwRy&sig=ZWpVmVJGUW4YY7q9Aqz71pkJcpl&hl=tr&sa=X&ved=0ahUKEwiVtOOIkq_YAhWILVAKHUB-Dx0Q6AEILTA#v=onepage&q=Comparing%20the%20psycholinguistic%20and%20the%20communicative%20paradigm%20of%20language%20proficiency.&f=false, Retrieved on 15 November, 2017.]
- De Jong, J.H.A.L., & Zheng, Y. (2011). *Research Note: Applying EALTA guidelines: A practical case study on Pearson Test of English Academic*. London: GB Pearson.
- DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2015). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability*, 28(3), 251-272.
- Demir, D., & Genç, A. (2016). The analysis of foreign language proficiency exam for the state employees based on the CEFR. *Hacettepe University Journal of Education*, 31(1), 53-65.
- Demirer, D. K. (2015). Reproduction of inequality through private out-of-school education. In: Carmo, M. (Ed.) *Education Applications and Development: Advances in Education and Educational Trends*. World Institute for Advanced Research and Science (WIARS), Lisbon: The Science Press.
- Dolgunsöz, E. (2016). A sudden change in Turkish education system: Public attitude towards dersane debates in Turkey. *E-International Journal of Educational Research (E-IJER)*, 7(2), 56-75.
- Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford: Oxford University Press.

- Eaton, S. E. (2010). *Formal, non-formal and informal learning: The case of literacy, essential skills, and language learning in Canada*. Calgary: Eaton International Consulting Inc.
- Eckes, T. (2009). *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: Section H: Many-Facet Rasch Measurement*. Strasbourg: Language Policy Division.
- Elliott, N., & Jordan, J. (2010). Practical strategies to avoid the pitfalls in grounded theory research. *Nurse Researcher*, 17(4), 29-40.
- Ellis, R. (1993). Second language acquisition and the structural syllabus. *TESOL Quarterly*, 27, 91-113.
- Erickson, G., & Figueras, N. (2010). *EALTA guidelines for good practice in language testing and assessment: Large scale dissemination days*. [On-line: http://www.ealta.eu.org/documents/archive/GGP_dissemination_report.pdf, Retrieved on 15 July, 2016.]
- Erkut, S., Alarcón, O., Coll, C. G., Tropp, L. R., & García, H. A. V. (1999). The dual-focus approach to creating bilingual measures. *Journal of Cross-Cultural Psychology*, 30(2), 206-218.
- European Association for Language Testing and Assessment (EALTA). (2006). *EALTA guidelines for good practice in language testing and assessment*. [On-line: <http://www.ealta.eu.org/documents/archive/guidelines/English.pdf>, Retrieved on 19 October, 2016.]
- European Communities. (2009). *Europass website*. [On-line: <http://europass.cedefop.europa.eu/>, Retrieved on 2 April, 2017.]
- Figueras, N. (2007). The CEFR, a lever for the improvement of language professionals in Europe. *The Modern Language Journal (MLJ)*, 91(4), 673-675.
- Finch, A. E. (2009). Europass and the CEFR: Implications for language teaching in Korea. *English Language & Literature Teaching*, 15(2), 1-23.
- Fordham, P. E. (1993). *Informal, non-formal and formal education programmes in YMCA George Williams College ICE301 Lifelong Learning Unit 2*. London, UK: YMCA George Williams College.
- Fulcher, G., Davidson, F., & Kemp, J. (2010). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, 28(1), 5-29.
- Girard, D., & Trim, J. (1988). *Learning and teaching modern languages for communication: Project no. 12, Final report of the project group (activities 1982-87)*. Strasbourg: Council for Cultural Cooperation.
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. New York, NY: Aldine De Gruyter.

- Glover, P. (2011). Using CEFR level descriptors to raise university students' awareness of their speaking skills. *Language Awareness*, 20(2), 121-133.
- Graddol, D. (2006). *English Next: Why global English may mean the end of "English as a foreign language"*. The United Kingdom: The British Council.
- Green, A. (2017). Linking tests of English for academic purposes to the CEFR: The score user's perspective. *Language Assessment Quarterly*, 0(0), 1-16.
- Halbherr, T., Schlienger, C., & Piendl, T. (2014). *Assessments for a digital world*. Paper presented at the annual AEA- Europe Tallinn Conference: Assessment of students in a 21st century world, 6-8 November, 2014, Tallinn, Estonia. [On-line: <https://www.research-collection.ethz.ch/handle/20.500.11850/195565>, Retrieved on 18 September 2017.]
- Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, 32(3), 317-336.
- Hasselgreen, A. (2005). Assessing the language of young learners. *Language Testing*, 22(3), 337-354.
- Holliday, A. (1992). Tissue rejection and informal orders in ELT projects: Collecting the right information. *Applied Linguistics*, 13(4), 403-424.
- Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, 103(2), 219-230.
- Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal (MLJ)*, 91(4), 663-667.
- Hulstijn, J. H. (2011). Language proficiency in native and non-native speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, 8(3), 229-249.
- Hulstijn, J. H., Schoonen, R., de Jong, N. H., Steinel, M. P., & Florjin, A. (2011). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the common European framework of reference for languages (CEFR). *Language Testing*, 29(2), 203-221.
- Ilic, G., & Stopar, A. (2015). Validating the Slovenian national alignment to CEFR: The case of the B2 reading comprehension examination in English. *Language Testing*, 32(4), 443-462.
- Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, 25, 385-402.
- International Language Testing Association (ILTA). (2000). *Code of Ethics*. [On-line: http://c.ymcdn.com/sites/www.iltaonline.com/resource/resmgr/docs/ilta_code_english.pdf, Retrieved on 20 September 2016.]

- International Language Testing Association (ILTA). (2007). *Guidelines for practice in English*. [On-line: http://c.ymcdn.com/sites/iltaonline.site-ym.com/resource/resmgr/docs/ilta_guidelines.pdf, Retrieved on 20 September 2016.]
- International Language Testing Association (ILTA). (2008). *Constitution for the International Language Testing Association*. [On-line: [http://c.ymcdn.com/sites/iltaonline.site-ym.com/resource/resmgr/docs/ilta_constitution\(1\).pdf](http://c.ymcdn.com/sites/iltaonline.site-ym.com/resource/resmgr/docs/ilta_constitution(1).pdf), Retrieved on 20 September 2016.]
- Isaacs, T., & Trofimovich, P. (Eds.). (2017). *Second language pronunciation assessment: Interdisciplinary perspectives*. Bristol: Multilingual Matters.
- Jarvis P. (1987). *Adult learning in the social context*. New York, NY: Routledge.
- Jones, N. (2007). Assessment and the national languages strategy. *Cambridge Journal of Education*, 37, 17-33.
- Jones, N. (2009). A comparative approach to constructing a multilingual proficiency framework: Constraining the role of standard-setting. In Figueras, N., & Noijons, J. (Eds.). *Linking to the CEFR levels: Research perspectives*, 35–44, Arnhem: Cito, EALTA.
- Jones, N., & Saville, N. (2009). European language policy: Assessment, learning, and the CEFR. *Annual Review of Applied Linguistics*, 29, 51-63.
- Jones, N., & Saville, N. (2014). *Learning oriented assessment: A systemic approach (Studies in Language Testing)*. Cambridge: Cambridge University Press.
- Johnson, R. B., & Christensen, L. B. (2012). *Educational research: Quantitative, qualitative, and mixed method approaches* (6th edition). Los Angeles: Sage.
- Kachru, B. B. (1992). World Englishes: Approaches, issues and resources. *Language Teaching*, 25(1), 1-14.
- Kaftandjieva, F. (2004). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the CEF: Section B: Standard setting*. Strasbourg: Language Policy Division.
- Karaboğa, K. (2013, 23 December). *İngilizce'ye yılda 100 milyar* [100 billion Turkish liras for learning English]. *Dünya* [The World]. [On-line: <https://www.dunya.com/ekonomi/ingilizceye-yilda-100-milyon-haberi-231840>, Retrieved on 10 November, 2017.]
- Kavaklı, N. (2017a). *Towards a continuum of professional development: Enhancing prospective EFL teachers' assessment literacy*. Paper presented at the XIII. European Conference on Social and Behavioral Sciences, 19- 22 May, 2017, Sofia, Bulgaria. [On-line: http://iassr2.org/rs/13_ab.pdf, Retrieved on 25 August, 2017.]
- Kavaklı, N. (2017b). *Utilizing European guidelines for establishing quality standards in language testing and assessment*. Paper presented at GlobELT 2017: An

International Conference on Teaching and Learning English as an Additional Language, 18- 21 May, 2017, Izmir, Turkey. [On-line: http://www.globeltconference.com/files/GlobELT2017_Conference_Book.pdf, Retrieved on 25 August, 2017.]

- Kavaklı, N., & Arslan, S. (2017). Applying EALTA Guidelines as baseline for the foreign language proficiency test in Turkey: The case of YDS. *International Journal of Curriculum and Instruction (IJCI)*, 9(1), 104-118.
- Kimura, Y., Nakata, Y., Ikeno, O., Naganuma, N., & Andrews, S. (2017). Developing classroom language assessment benchmarks for Japanese teachers of English as a foreign language. *Language Testing in Asia*, 7(3), 1-14.
- Külekçi, E. (2016). A concise analysis of the Foreign Language Examination (YDS) in Turkey and its possible washback effects. *International Online Journal of Education and Teaching (IOJET)*, 3(4), 303-315.
- La Belle, T. J. (1982). Formal, non-formal and informal education: A holistic perspective on lifelong learning. *International Review of Education*, 28(2), 159-175.
- Lambert, D., & Lines, D. (2000). *Understanding assessment: Purposes, perceptions, practice*. London, UK: Routledge Falmer.
- Levy, M. (1997). *Computer-assisted language learning: Context and conceptualization*. Oxford: Oxford University Press.
- Lightbown, P., & Spada, N. (1990). Focus on form and corrective feedback in communicative language teaching: Effects on second language learning. *Studies in Second Language Acquisition*, 12, 429-448.
- Little, D. (2005). The common European framework and the European language portfolio: Involving learners and their judgements in the assessment process. *Language Testing*, 22(3), 321-336.
- Little, D. (2007). The common European framework of reference for languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal (MLJ)*, 91, 645-655.
- Malone, M. E. (2017). Training in language assessment. In Shohamy et al. (Eds.), *Language Testing and Assessment, Encyclopedia of language and education* (3rd edition), 225-239. Cham: Springer.
- Marczyk, G., DeMatteo, D., & Festinger, D. (2005). *Essentials of research design and methodology*. New Jersey: John Wiles and Sons Inc.
- Marín, G., & Marín, B. V. (1991). *Research with Hispanic populations*. Newbury Park, CA: Sage.
- Martyniuk, W. (2010). *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*. Cambridge: Cambridge University Press.

- Martyniuk, W., & Noijons, J. (2007). *Executive summary of results of a survey on the use of the CEFR at national level in the Council of Europe Member States*. Document presented at the Council of Europe Intergovernmental Language Policy Forum: The Common European Framework of Reference for Languages (CEFR) and the Development of Language Policies: Challenges and Responsibilities, 6-8 February, 2007, Strasbourg, France. [On-line: http://www.coe.int/t/dg4/linguistic/Source/Survey_CEFR_2007_EN.doc, Retrieved on 24 February 2017.]
- Mendoza, A. A. L., & Arandia, R. B. (2009). Language testing in Colombia: A call for more teacher education and teacher training in language assessment. *Issues in Teachers' Professional Development*, 11(2), 55-70.
- Menken, K. (2008). *English learners left behind: Standardized testing as language policy*. Clevedon: Multilingual Matters.
- Mills, G. E. (2003). *Action research a guide for the teacher researcher*. Boston: Pearson Education.
- Mirici, İ. H. (2008). Development and validation process of a European language portfolio model for young learners. *Turkish Online Journal of Distance Education (TOJDE)*, 9(2), 26-34.
- Mirici, İ. H. (2015). Contemporary ELT practices across Europe. *International Journal of Language Academy (IJLA)*, 3(4), 1-8.
- Mirici, İ. H., & Kavaklı, N. (2017). Teaching the CEFR-oriented practices effectively in the MA program of an ELT department in Turkey. *International Online Journal of Education and Teaching (IOJET)*, 4(1), 74-85.
- Morrow, L. M. (2012). *Literacy development in the early years: helping children read and write* (7th edition). Boston: Pearson.
- North, B. (2005). *Le Cadre européen commun de référence: Introduction*. [The Common European framework of reference: Introduction]. Paper presented at the Journée Pédagogique, 15 June, 2005, Paris, France. On-line: <http://www.alliance-us.org/dg/documentupload/cecrBrianNorth.pdf>, Retrieved on 9 April 2017.]
- North, B. (2014). *The CEFR in practice*. Cambridge: Cambridge University Press.
- North, B., Martyniuk, W., & Panthier, J. (2010). Introduction: The manual for relation language examinations to the common European framework of reference for languages in the context of the Council of Europe's work on language education. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*, 1–17. Cambridge: Cambridge University Press.
- Nunan, D. (1992). *Research methods in language learning*. Cambridge: Cambridge University Press.

- Öğrenci Seçme ve Yerleştirme Merkezi (ÖSYM). (2013a). *Basın duyurusu*. [Press releasement]. [On-line: <http://osym.gov.tr/belge/1-14909/basin-duyurusu-2013-yabanci-dil-bilgisi-seviye-belirlem-.html>, Retrieved on 25 April, 2017.]
- Öğrenci Seçme ve Yerleştirme Merkezi (ÖSYM). (2016a). *2016 yabancı dil bilgisi seviye tespit sınavı (YDS) kılavuzu* [A guide for foreign language proficiency examination, 2016]. Ankara: ÖSYM. [On-line: <http://dokuman.osym.gov.tr/pdfdokuman/2016/YDSILKBAHAR/2016YDSilkbhrklvz02022016.pdf>, Retrieved on 28 May, 2017.]
- Öğrenci Seçme ve Yerleştirme Merkezi (ÖSYM). (2016b). *Yabancı dil sınavlarının eşdeğerliklerini belirleme yönergesi* [The directions of defining equivalences for foreign language proficiency examinations]. [On-line: <http://dokuman.osym.gov.tr/pdfdokuman/2016/GENEL/EsdegerlikYonergesi26022016.pdf>, Retrieved on 14 May, 2017.]
- Öğrenci Seçme ve Yerleştirme Merkezi (ÖSYM). (2016c). *Yabancı dil sınavları eşdeğerlikleri* [The equivalences of foreign language proficiency examinations]. [On-line: <http://www.yok.gov.tr/documents/10279/30814109/EsdegerlikTablosu.pdf/>, Retrieved on 14 May, 2016.]
- Özel Öğretim Kurslar, Dershaneler ve Etüt Eğitim Merkezleri Birliği Derneği. (ÖZ-KUR-DER). (2011). *Kamuoyuna açıklama*. [Declaration to the Public]. The Association of Private Educational Establishments and Study Centers of Turkey. [On-line: http://www.ozkurder.com/bilgilendirme/kamuya_bilgi.htm., Retrieved 2 May, 2017.]
- Özel Öğretim Kurslar, Dershaneler ve Etüt Eğitim Merkezleri Birliği Derneği. (ÖZ-KUR-DER). (2017). *Çeşitli kursların sorunları*. [Problems of some courses]. The Association of Private Educational Establishments and Study Centers of Turkey. [On-line: http://ozkurder.com/wp-content/uploads/2017/01/cesitli_kurslarin_sorunlari.pdf, Retrieved 2 August, 2017.]
- Palomba, C. A., & Banta, T. W. (1999). *Assessment essentials: Planning, implementing, and improving assessment in higher education*. San Francisco: Jossey-Bass Inc.
- Pearson English. (2014). *English: The world's language (infographic)*. Pearson. [On-line: https://www.english.com/english_learning_infographic, Retrieved on 17 September, 2016.]
- Piccardo, E., North, B., & Maldina, E. (2017). *QualiCEFR: A quality assurance template to achieve innovation and reform in language education through CEFR implementation*. Proceedings of the ALTE 6th International Conference on Learning and Assessment: Making the Connections, 3-5 May, Bologna, Italy. [On-line: <http://events.cambridgeenglish.org/alte2017-test/perch/resources/alte-2017-proceedings-final.pdf>, Retrieved on 22 August, 2017.]

- Platton, M. Q. (2015). *Qualitative research and evaluation methods: Integrating theory and practice* (4th edition). Thousand Oaks, CA: Sage.
- Purpura, J. E. (2017). Assessing meaning. In: Shohamy E., or I., May S. (Eds.) *Language testing and assessment. Encyclopedia of Language and Education* (3rd edition), 63-76. Cham: Springer.
- Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies*, 7(2), 105-125.
- Rehorick, S., & Lafargue, C. (2005). *The European language portfolio and its potential for Canada*. Report on the proceedings of the National workshop on language portfolios, 12- 14 October, 2005. University of New Brunswick, Edmonton, Alberta, Canada. [On-line: http://www.unbf.ca/L2/esources/PDFs/ELP/UNB_ELP_fullreport.pdf, Retrieved on 6 September 2017.]
- Richard-Amato, P. (1988). *Making it happen: Interaction in the second language classroom*. New York, NY: Longman.
- Roca-Varela, M. L., & Palacios, I. M. (2013). How are spoken skills assessed in proficiency tests of general English as a foreign language? A preliminary survey. *International Journal of English Studies (IJES)*, 13(2), 53-68.
- Romi, S., & Schmida, M. (2009). Non-formal education: A major educational force in the postmodern era. *Cambridge Journal of Education*, 39(2), 257-273.
- San, İ. (2016). Assessment for learning: Turkey case. *Universal Journal of Educational Research*, 4(1), 137-143.
- Sarantakos, S. (2005). *Social Research* (3rd edition). Basingstoke: Palgrave Macmillan.
- Sarıçoban, A. (2011). A study on the English language teachers' preparation of tests. *Hacettepe University Journal of Education*, 41, 398-410.
- Saville, N. (2005). An interview with John Trim at 80. *Language Assessment Quarterly*, 2(4), 263-288.
- Schaerer, R. (2005). *European language portfolio: Interim report 2005 with executive summary*. Strasbourg: Language Policy Division.
- Scheerens, J., Glas, C., & Thomas, S. M. (2003). *Educational evaluation, assessment and monitoring: A systemic approach*. New York, NY: Taylor and Francis.
- Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing*, 18(4), 373-391.
- Silova, I., Budiene, V., & Bray, M. (Eds.). (2006). *Education in a hidden marketplace: Monitoring of private tutoring*. New York, NY: Open Society Institute.
- Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, 17, 38-62.

- Sleeter, C. E., & Carmona, J. F. (2017). *Un-standardizing curriculum: Multicultural teaching in the standards-based classroom* (2nd edition). New York, NY: Teachers College Press.
- Southgate, D. (2009). *Determinants of shadow education: A cross-national analysis*. (Unpublished Doctoral Dissertation). The Ohio State University, Ohio, The USA.
- Spinelli, C. G. (2007). Addressing the issue of cultural and linguistic diversity and assessment: Informal evaluation measures for English language learners. *Reading and Writing Quarterly*, 24(1), 101-118.
- Spöttl, C., Kremmel, B., Holzknicht, F., & Alderson, J. C. (2016). Evaluating the achievements and challenges in reforming a national language exam: The reform team's perspective. *Papers in Language Testing and Assessment*, 5(1), 1-22.
- Stoynoff, S. (2012). Looking back and forward at classroom-based language assessment. *ELT Journal*, 66(4), 523-532.
- Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics*, 16, 371-391.
- Taber, K. S. (2000). Case studies and generalizability: Grounded theory and research in science education. *International Journal of Science Education*, 22, 469-87.
- Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English-language test scores onto the common European framework of reference: An application of standard setting methodology (TOEFL iBT Series Report No. 06)*. Princeton, NJ: Educational Testing Service.
- Taylor, L. (2006). The changing landscape of English: implications for language assessment. *ELT Journal*, 60(1), 51-60.
- Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes: defining and operationalizing the test construct. *Journal of English for Academic Purposes*, 10(2), 89,101.
- Toncheva, S., Zlateva, D., & John, P. (2017). *Developing an assessment methodology for a universal maritime English proficiency test for deck officers*. Paper presented at the 18th Annual General Assembly of the International Association of Maritime Universities (IAMU), 11-13 October, 2017, Varna, Bulgaria. [On-line: https://www.researchgate.net/publication/320419205_Developing_an_assessment_methodology_for_a_universal_Maritime_English_proficiency_test_for_deck_officers, Retrieved on 20 October, 2017.]
- Trim, J. L. M. (2001). *The work of the Council of Europe in the field of modern languages, 1957-2001*. Paper presented on the European Day of Languages, 26 September, 2001, European Centre for Modern Languages, Graz, Austria.

[On-line:

http://www.ecml.at/Portals/1/resources/John%20Trim%20collection/Trim_TheWorkOfTheCouncilOfEurope_ModernLanguages_1957_2001.pdf, Retrieved on 25 September, 2017.]

Trim, J. L. M. (2005). *The role of the Common European Framework of Reference for Languages in teacher training*. Lecture delivered during the Ceremony of the 10th Anniversary of the European Centre for Modern Languages of the Council of Europe, 16 September, 2005, Graz, Austria. [On-line: [http://www.europarl.europa.eu/RegData/etudes/etudes/join/2013/495871/IP-OL-CULT_ET\(2013\)495871_EN.pdf](http://www.europarl.europa.eu/RegData/etudes/etudes/join/2013/495871/IP-OL-CULT_ET(2013)495871_EN.pdf), Retrieved on 17 November, 2016.]

United Nations (UN). (1990). *Convention on the Rights of the Child*. Human Rights Office of the High Commissioner, the United Nations. [On-line: <http://www.ohchr.org/EN/ProfessionalInterest/Pages/CRC.aspx>, Retrieved on 20 November, 2016.]

United Nations Educational, Scientific and Cultural Organization (UNESCO). (1972). *Learning to be: The world of education today and tomorrow*. Paris: UNESCO.

Valax, P. (2011). *The Common European Framework of Reference for Languages: A critical analysis of its impact on a sample of teachers and curricula within and beyond Europe*. (Unpublished Doctoral Dissertation). University of Waikato, Hamilton, New Zealand.

van Ek, J. A., & Trim, J. L. M. (1990). *Threshold*. Cambridge: Cambridge University Press.

Van Nijlen, D., & Janssen, R. (2014). *Measuring 21st century skills through national assessments: The case of information processing skills*. Paper presented AEA- Europe Tallinn Conference: Assessment of students in a 21st century world, 6-8 November, 2014, Tallinn, Estonia. [On-line: <https://lirias.kuleuven.be/handle/123456789/471002>, Retrieved on 20 October, 2017.]

Verhelst, N. (2004a). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the CEF: section C: Classical test theory*. Strasbourg: Language Policy Division.

Verhelst, N. (2004b). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the CEF: section E: Generalizability theory*. Strasbourg: Language Policy Division.

Verhelst, N. (2004c). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the CEF: section F: Factor analysis*. Strasbourg: Language Policy Division.

Verhelst, N. (2004d). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the CEF: Section G: Item response theory*. Strasbourg: Language Policy Division.

- Walvoord, B. E., & Anderson, V. J. (2010). *Effective grading: A tool for learning and assessment in college* (2nd edition). San Francisco: Jossey-Bass Inc.
- Wang, H.-P., Kuo, B.-C., Tsai, Y.-H., & Liao, C.-H. (2012). A CEFR-based computerized adaptive testing system for Chinese proficiency. *The Turkish Online Journal of Educational Technology (TOJET)*, 11(4), 1-12.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.
- Wilkins, D. (1973). *The linguistic and situational content of the common core unit/credit system*. Strasbourg: Council of Europe.
- Wools, S. (2015). *All about validity: An evaluation system for the quality of educational assessment*. Enschede: University of Twente.
- Wu, R. W. (2008). An investigation of the relationships between strategy use and GEPT test performance. *English Teaching & Learning*, 32, 35–69.
- Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, 27(2), 147-170.
- Zumbo, B. D. (2015). *Consequences, side effects and the ecology of testing: Keys to considering assessment in 'In Vivo'*. Keynote address at the annual meeting of the Association for Educational Assessment - Europe (AEA-Europe), 5-7 November, 2015, Glasgow, Scotland. [On-line: https://www.researchgate.net/profile/Bruno_Zumbo/publication/297277410_Consequences_Side_Effects_and_the_Ecology_of_Testing_Keys_to_Considering_Assessment_'In_Vivo'_Keynote_address/links/56df741d08ae9b93f79a9864/Consequences-Side-Effects-and-the-Ecology-of-Testing-Keys-to-Considering-Assessment-In-Vivo-Keynote-address.pdf, Retrieved on 12 May, 2017.]

APPENDIX-A: Questionnaire on the European Standards for Establishing Quality Profiles in Language Examinations

(17.12.2016)

Dear Participant(s),

This questionnaire, as a part of research study, is designed to uncover the testing and assessment practices of English language schools in Turkey within the scope of some European standards. From that point of view, some European standards for establishing quality profiles in exams are listed below with the help of international guidelines proposed by Association of Language Testers in Europe (ALTE), European Association for Language Testing and Assessment (EALTA), International Language Testing Association (ILTA) and The Association for Educational Assessment- Europe (AEA-Europe). You are invited to contribute to the findings of a Ph.D. dissertation as your help is highly appreciated and important to complete this study.

To elaborate briefly, the questionnaire is composed of 87 Likert-type response items, each with five (5) options. These numbers are the indicators of *Strongly Disagree* (1), *Disagree* (2), *Not sure* (3), *Agree* (4) and *Strongly Agree* (5) respectively. At that point, you are expected to read each statement carefully and circle the number that **BEST** describes your opinion. Please keep in mind that there is not any **CORRECT** or **FALSE** answer. By the way, any information that can identify you will remain confidential. The information in this study will be used only for research purposes and in ways that will not reveal who you are. You may not benefit from being a part of this study but your participation may help to improve the CEFR oriented testing and assessment practices of English language schools in Turkey. Thank you in advance for your efforts and invaluable contribution to the findings of this study.

WHO YOU ARE...

1. Gender: Female Male

2. Age: 18-25 years
..... 26-35 years
..... 36-45 years
..... 46-55 years
..... 56-65 years

- 3. Years of experience:** less than 5 years
 5 to 9 years
 10 to 14 years
 more than 14 years

- 4. Occupational field:** Teacher
 (You can mark more than one option) Administrator
 Test (-item) developer
 (If) other, please specify

European Standards for Establishing Quality Profiles in Exams		Strongly Disagree	Disagree	Not Sure	Agree	Strongly Agree
1	The tests are based on a theoretical construct or a model (e.g. communicative competence).	1	2	3	4	5
2	The purpose, context of use, and target population for the tests are appropriately stated.	1	2	3	4	5
3	The tests cover the full range of knowledge and skills relevant and useful to real world situations and authentic language use.	1	2	3	4	5
4	The test scores correlate with a recognized external criterion which measures the same area of knowledge or ability (e.g. the CEFR).	1	2	3	4	5
5	Criteria for selection and training of test constructors and expert judgment are involved both in test construction, and in the review and revision of the tests.	1	2	3	4	5
6	The tests are comparable with parallel examinations across different administrations in terms of content, consistency and grade boundaries.	1	2	3	4	5
7	Evidence of the tests' linkage to an external reference system (e.g. the CEFR) is available through alignment chart.	1	2	3	4	5
8	The purpose of the tests is clearly defined.	1	2	3	4	5
9	The content of the tests is consistent with the stated goal for which the test is being administered.	1	2	3	4	5
10	Discriminant validity sub-scores are supported by means of logical and empirical evidences.	1	2	3	4	5
11	The test takers' characteristics are clearly defined.	1	2	3	4	5

12	The tests are appropriate to the overall abilities of the test-takers.	1	2	3	4	5
13	The tests have been previously tried out on a sample of persons from the same general population as the target test-takers.	1	2	3	4	5
14	The test results are reliable enough to make accurate decisions.	1	2	3	4	5
15	The degree of reliability of the test is demonstrated by numerical data.	1	2	3	4	5
16	The format of the tests is suitable, and its contextual use is clearly defined.	1	2	3	4	5
17	The test takers are familiar with the actual test format(s).	1	2	3	4	5
18	The format and features of the tests can be fairly applied in the real testing situations.	1	2	3	4	5
19	The tests are relevant to the proposed test population and/or to the test item domain.	1	2	3	4	5
20	The proposed test population/content resemble the developmental sample closely.	1	2	3	4	5
21	It is easy to produce equivalent or equated forms of the tests being used.	1	2	3	4	5
22	It is easy to score the tests, report the test scores and interpret the results.	1	2	3	4	5
23	The tests require a great deal of training before they are conducted.	1	2	3	4	5
24	It costs a lot to procure and administer the tests.	1	2	3	4	5
25	It costs a lot to score the tests.	1	2	3	4	5
26	The tests are readily available.	1	2	3	4	5
27	The tests are societally and institutionally acceptable.	1	2	3	4	5
28	The tests are acceptable in the eyes of teachers, parents and administrators.	1	2	3	4	5
29	All centers are selected to administer the tests according to clear, transparent, established procedures, and have access to regulations about how to do so.	1	2	3	4	5
30	Examination papers are delivered in excellent condition, and by secure means to the scoring centers.	1	2	3	4	5
31	The examination administration system has appropriate support systems (e.g. phone hotline, web services etc.).	1	2	3	4	5
32	The results are adequately protected by the security, and confidentiality of the results and certificates is enabled.	1	2	3	4	5
33	The examination system provides support for candidates with special needs.	1	2	3	4	5

34	Marking is sufficiently accurate and reliable for purpose and type of the test.	1	2	3	4	5
35	How marking is carried out is documented and explained through raters' reliability estimates.	1	2	3	4	5
36	The data is collected on an adequate and representative sample of candidates, and not influenced by factors like L1, country of origin, gender, age and ethnic origin.	1	2	3	4	5
37	Item-level data (e.g. for computing the difficulty, discrimination, reliability and standard errors of measurement of the examination) is collected from an adequate sample of candidates.	1	2	3	4	5
38	The test administration system communicates the test results to candidates, and if required, to examination centers (e.g. schools) promptly and clearly.	1	2	3	4	5
39	The stakeholders are informed on the context, purpose, use of the tests, and the overall reliability of the test results appropriately.	1	2	3	4	5
40	Stakeholders are informed about how to interpret and use the test results appropriately.	1	2	3	4	5
41	The test takers are supplied with different response items (e.g. short answer, sentence correction, gap filling, multiple choice).	1	2	3	4	5
42	The candidates are provided with non-item based task types (e.g. writing tasks, speaking tasks).	1	2	3	4	5
43	The marking scheme, rubrics, answer keys and rating scales are readily available.	1	2	3	4	5
44	The equivalence between different versions of the tests (e.g. year by year) are verified.	1	2	3	4	5
45	The actions to improve the quality of teaching and learning are taken after the implementation of each test.	1	2	3	4	5
46	The test items keep pace with changes in the current ELT curriculum.	1	2	3	4	5
47	There is a publicly available report on the linking process between tests in use and the Reference Supplement, such as the CEFR.	1	2	3	4	5
48	As a part of the linkage to the CEFR, the tests correspond to the procedures recommended in the Manual and Reference Supplement.	1	2	3	4	5
49	Test specifications and tasks are spelled out in detail.	1	2	3	4	5
50	The tasks and test items are edited before (pre)testing.	1	2	3	4	5
51	The test materials are kept in a safe place.	1	2	3	4	5
52	Scoring procedures are carefully followed.	1	2	3	4	5
53	Items written by non-native speakers of the target language are checked by someone with a high-level of competence in the target language.	1	2	3	4	5

54	Test takers are treated with courtesy and respect during the testing process.	1	2	3	4	5
55	Test takers read or listen to descriptive information and test instructions in advance of testing.	1	2	3	4	5
56	Test takers are well aware of the consequences of not taking the test.	1	2	3	4	5
57	Test takers can inform appropriate person(s), who are specified by the organization to be responsible for testing, if they believe that testing conditions have affected their results.	1	2	3	4	5
58	Test item writers are trained before test administration.	1	2	3	4	5
59	Overall evaluation of the total program, and assessment of educational systems are taken into consideration in testing procedures.	1	2	3	4	5
60	Innovative assessment techniques are taken into consideration while designing tests.	1	2	3	4	5
61	European perspective to the world-wide interest in assessment is adopted.	1	2	3	4	5
62	Establishing standards as a way of disseminating quality in assessment is the core element in testing and assessment practices.	1	2	3	4	5
63	The tests in use support different cultural and educational contexts.	1	2	3	4	5
64	The test takers' place in the assessment process is well-defined.	1	2	3	4	5
65	What is good for the individual in assessment aligns with the United Nations Convention on the Rights of the Child.	1	2	3	4	5
66	Ethical considerations are given prominence in assessment procedures.	1	2	3	4	5
67	The assessment belongs to the rights of the test takers; not to those who devise and administer the tests.	1	2	3	4	5
68	The cornerstones of assessment (e.g. validity, practicality, impact on stakeholders) are carefully addressed.	1	2	3	4	5
69	The results in the light of the essential quality aspects are meaningful and useful.	1	2	3	4	5
70	Assessment translates the evidence that the results are defensible in different educational settings for further use.	1	2	3	4	5
71	The purpose of the assessment supports the overall education of test takers.	1	2	3	4	5
72	The assessment bases its rationale on the intended learning, which underlies a particular educational process.	1	2	3	4	5
73	Assessment procedures provide information that confirms the aims of the Common European Framework of Reference for Languages.	1	2	3	4	5

74	The kinds of assessment allow for feedback on the performance of the on-going educational system.	1	2	3	4	5
75	Decision makers have the opportunity to evaluate programs and allocate resources by means of test results.	1	2	3	4	5
76	The core elements of the Common European Framework of Reference for Languages are distinguished which follow the assessment development cycle.	1	2	3	4	5
77	Possible evidences are presented to check whether the standard requirements are met by the test administered.	1	2	3	4	5
78	The assessment applied in the institution/organization covers standardized tests.	1	2	3	4	5
79	The assessment applied in the institution/organization covers school-based (summative) examinations.	1	2	3	4	5
80	The assessment applied in the institution/organization covers vocational (performance) assessment.	1	2	3	4	5
81	The assessment applied in the institution/organization covers learning outcomes of a curriculum (formative assessment).	1	2	3	4	5
82	The assessment applied in the institution/organization covers competency tests.	1	2	3	4	5
83	The tests are piloted before they are administered to the target population.	1	2	3	4	5
84	Test results are scored via automated scoring machines.	1	2	3	4	5
85	Test results are scored via human scoring.	1	2	3	4	5
86	Test takers are provided with contemporary self-assessment tools such as the European Language Portfolio (ELP).	1	2	3	4	5
87	Traditional assessment practices are in use for test takers.	1	2	3	4	5

THANK YOU!!!

APPENDIX-B: Semi-Structured Interview Forms conducted with the Directors of Selected Private Institutions and that of ÖZ-KUR-DER (Original in Turkish)

Kurs:

Şube Sayısı:

Öğretmen Sayısı:

Yaklaşık Öğrenci Sayısı:

Diğer Sayısal Veriler:

Mülakat Soruları

- (1) Kurum içinde yürütülen ölçme-değerlendirme hakkında lütfen kısaca bilgi veriniz.
- (2) Bu çalışmalar, herhangi bir Avrupa standardına uygunluk gösteriyor mu? Cevabınız evet ise, bunlar nelerdir?
- (3) Kurum içinde kullanılan ölçme-değerlendirme kriterleri ve araçları nelerdir?
- (4) Kurum içi ölçme-değerlendirme uygulamalarında karşılaşılan güçlük ve sorunlar nelerdir?
- (5) Kurum içi ölçme-değerlendirme uygulamalarının iyileştirilmesi için önerileriniz nelerdir?
- (6) Tatbiki ölçme ve değerlendirme uygulamalarının ülke genelinde iyileştirilmesi için önerileriniz nelerdir?

Katkılarınız ve katılımınız için teşekkür ederiz.

APPENDIX-C: Semi-Structured Interview Forms conducted with the Directors of Selected Private Institutions and that of ÖZ-KUR-DER (Translated into English)

Name of the English Language School:

The number of Branches:

The number of Teachers:

Approximate number of Students:

Other Numeric Data:

Semi-Structured Interview Questions

(1) Please provide some information on the testing and assessment practices conducted within your institution.

(2) Are these practices aligned with any European standards? If yes, please provide some information about those standards.

(3) Please provide some information about the instruments and the criteria set for testing and assessment practices.

(4) Please provide some information on the difficulties and problems mostly encountered in testing and assessment practices.

(5) Please provide some recommendations in order to enhance the on-going testing and assessment practices within your institution.

(6) Please provide some recommendations in order to enhance the on-going testing and assessment practices across the country.

Thank you for your dearest concern and participation in this study.

APPENDIX-D: Ethics Committee Approval

Form: 40

Tez Çalışması Etik Kurul İzin Muafiyeti Formu

01 / 06 / 2015

Hacettepe Üniversitesi
Eğitim Bilimleri Enstitüsü
Yabancı Diller Eğitimi Anabilim Dalı Başkanlığı'na

Tez Başlığı / Konusu:	CEFR Oriented Testing and Assessment Practices in Non-formal English Language Schools in Turkey
------------------------------	---

Yukarıda başlığı/konusu gösterilen tez çalışmam:

1. İnsan ve hayvan üzerinde deney niteliği taşımamaktadır,
2. Biyolojik materyal (kan, idrar vb. biyolojik sıvılar ve numuneler) kullanılmasını gerektirmemektedir.
3. Beden bütünlüğüne müdahale içermemektedir.
4. Gözlemsel ve betimsel araştırma (anket, ölçek/skala çalışmaları, dosya taramaları, veri kaynakları taraması, sistem-model geliştirme çalışmaları) niteliğinde değildir.

Hacettepe Üniversitesi Etik Kurullar ve Komisyonlarının Yönergelerini inceledim ve bunlara göre tez çalışmamın yürütülebilmesi için herhangi bir Etik Kuruldan izin alınmasına gerek olmadığını; aksi durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

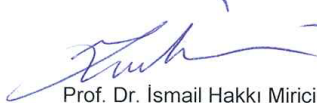

Nurdan Kavaklı
(Öğrencinin Adı Soyadı, İmzası)

Öğrenci Bilgileri

Adı Soyadı	Nurdan Kavaklı
Öğrenci No	N11264117
Anabilim Dalı	Yabancı Diller Eğitimi
Programı	İngiliz Dili Eğitimi
Statüsü	<input type="checkbox"/> Yüksek Lisans <input type="checkbox"/> Doktora <input checked="" type="checkbox"/> Bütünleşik Dr.

Danışman Görüşü ve Onayı

Çalışma, herhangi biçimde özel haklar ve sorumluluklar, bireysel özel bilgiler, kişisel hak ihlali sayılacak bir konu ile ilgili unsurlar yer içermediğinden ve çalışmanın verileri etik açıdan bir sakınca doğuracak özellikte olmadığından etik kurul izini alınmasına gerek duyulmamıştır.


Prof. Dr. İsmail Hakkı Mirici
(İmza)
(Danışmanın Ünvanı, Adı ve Soyadı)

APPENDIX-E: Declaration of Ethical Conduct

I hereby declare that...

- I have prepared this thesis in accordance with the thesis writing guidelines of the Graduate School of Educational Sciences of Hacettepe University;
- all information and documents in the thesis/dissertation have been obtained in accordance with academic regulations;
- all audio visual and written information and results have been presented in compliance with scientific and ethical standards;
- in case of using other people's work, related studies have been cited in accordance with scientific and ethical standards;
- all cited studies have been fully and decently referenced and included in the list of References;
- I did not do any distortion and/or manipulation on the data set,
- and **NO** part of this work was presented as a part of any other thesis study at this or any other university.

30/01/2018



Nurdan KAVAKLI

APPENDIX-F: Dissertation Originality Report

30/01/2018

HACETTEPE UNIVERSITY
Graduate School of Educational Sciences
To the Department of Foreign Language Education

Thesis Title: CEFR ORIENTED TESTING AND ASSESSMENT PRACTICES IN NON-FORMAL ENGLISH LANGUAGE SCHOOLS IN TURKEY

The whole thesis that includes the *title page, introduction, main chapters, conclusions and bibliography section* is checked by using **Turnitin** plagiarism detection software take into the consideration requested filtering options. According to the originality report obtained data are as below.

Time Submitted	Page Count	Character Count	Date of Thesis Defence	Similarity Index	Submission ID
30/01/2018	237	483983	16/01/2018	7%	908788116

Filtering options applied:

1. Bibliography excluded
2. Quotes included
3. Match size up to 5 words excluded

I declare that I have carefully read Hacettepe University Graduate School of Educational Sciences Guidelines for Obtaining and Using Thesis Originality Reports; that according to the maximum similarity index values specified in the Guidelines, my thesis does not include any form of plagiarism; that in any future detection of possible infringement of the regulations I accept all legal responsibility; and that all the information I have provided is correct to the best of my knowledge.

I respectfully submit this for approval.

Name Lastname: Nurdan KAVAKLI
Student No.: N11264117
Department: Foreign Language Education
Program: English Language Teaching
Status: Masters Ph.D. Integrated Ph.D.

ADVISOR APPROVAL

APPROVED
Prof. Dr. İsmail Hakkı MİRİCİ

APPENDIX-G: Yayınlama ve Fikrî Mülkiyet Hakları Beyanı

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kâğıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversite'ye verilen kullanım hakları dışındaki bütün fikrî mülkiyet haklarım bende kalacak, tezimin tamamının veya bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinleri yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversite'ye teslim etmeyi taahhüt ederim.

Tezimin/Raporumun tamamı dünya çapında erişime açılabilir ve bir kısmı veya tamamının fotokopisi alınabilir.

(Bu seçenekle teziniz arama motorlarında indekslenebilecek, daha sonra tezinizin erişim statüsünün değiştirilmesini talep etmeniz ve kütüphane bu talebinizi yerine getirirse bile, teziniz arama motorlarının ön belleklerinde kalmaya devam edebilecektir)

Tezimin/Raporumun tarihine kadar erişime açılmasını ve fotokopi alınmasını (İç Kapak, Özet, İçindekiler ve Kaynakça hariç) istemiyorum.

(Bu sürenin sonunda uzatma için başvuruda bulunmadığım takdirde, tezimin/raporumun tamamı her yerden erişime açılabilir, kaynak gösterilmek şartıyla bir kısmı veya tamamının fotokopisi alınabilir).

Tezimin/Raporumun 11/02/2020 tarihine kadar erişime açılmasını istemiyorum ancak kaynak gösterilmek şartıyla bir kısmı veya tamamının fotokopisinin alınmasını onaylıyorum.

Serbest Seçenek/ Yazarın Seçimi:

.....
.....
.....

30/01/2018

Nurdan KAVAKLI

