Available online at ijci.wcci-international.org

International Journal of Curriculum and Instruction 11(1) (2019) 65–88

IJCI
International Journal of Curriculum and Instruction

State-of-the-art of language testing and assessment in non-formal education: The case of English language schools in Turkey

Nurdan Kavaklı [a] *, İsmail Hakkı Mirici [b]

[a] *Assist. Prof. Dr., Izmir Demokrasi University, Turkey*

[b] *Prof. Dr., Near East University, Northern Cyprus*

Abstract

Purveying insights from a mixed-method research design, this study aims to enlighten the exploitation of the European guidelines in language testing and assessment practices in non-formal educational settings. Accordingly, three non-formal English language schools renowned for quality in Turkey were taken to in-depth analysis in order to offer a general paradigm from a sample of leading professionals on the utilization of the European benchmarks in language testing and assessment practices. The results have yielded that (1) there is a need for a more practical curriculum molded with a real auditing system for the enhancement of the current language testing and assessment practices; (2) there is a request for the validation process for language certificate examinations implemented in non-formal educational settings; (3) there is a demand for cooperation amidst the allies for the standardization process in language testing and assessment practices. The results are laced with some recommendations and implications for language testing and assessment.

Keywords: Language testing; assessment; nonformal education; CEFR; EFL; ELT.

1. Introduction

At the 20th session of the Standing Conference of the Ministers of Education of the Council of Europe (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 (ELP) as a self-assessment tool across Europe (CoE, 2000). Correlatively, the Common European Framework of Reference for Languages (CEFR) was adopted (CoE, 2001). Thus, the use

* Nurdan Kavaklı. Tel.: +090-232-260-1001
E-mail address: nurdan.kavakli@idu.edu.tr

of the European standards was taken as the basic premise for good practice in foreign language education.

As this was not the case solely for English-speaking countries, albeit outside English-speaking countries, the CEFR was nestled as a fundamental basis for the reconsideration of language teaching, testing and assessment practices in Turkey, as well (MoNE, 2005). Therefore, the CEFR was exploited to underpin all of the assessment and certification credentials by Turkish Ministry of National Education (MoNE) (Mirici, 2015).

Besides, there has been an ongoing demand for learning English outside schools; through private educational institutions. Herein, non-formal educational settings are drawing interest; however, there is not a robust auditing system within these institutions compared to formal educational settings, and the language testing and assessment practices carried out within these institutions are somehow fuzzy (Kavaklı & Mirici, 2018). Additionally, the estimated number of those enrolling in private institutions to meet their further demands in learning is 71.8% by the Council of Higher Education (2007) in Turkey, and the revenue gathered is approximately 600-750 Turkish Liras (Karaboğa, 2013).

Herein, the overall aim of this research is to delve into the utilization of the European benchmarks in language testing and assessment practices conducted within English language schools, those listed as the non-formal private institutions in Turkey. Assumed to do so, how well they pursue the applications and basic principles designated by the CEFR, the criteria defined by the European Association for Language Testing and Assessment (henceforth EALTA), the guideline assigned by the International Language Testing Association (henceforth ILTA) and the standards set by the Association of Language Testers in Europe (henceforth ALTE) are explored.

2. The European Standards of Language Testing and Assessment: CEFR

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 postwar

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.

However, the birth of the Framework in company with that of the European Language Portfolio (ELP) was accepted as the Rüschlikon 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 Rüschlikon 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 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. 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 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 level specification 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, 2005, 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. 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 (CoE, 2009a) as 'A Manual for Relating Language Examinations to the CEFR'. 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. 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 (Alderson, 2007; Mirici & Kavaklı, 2017).

3. The European Standards of Language Testing and Assessment: EALTA

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, and the adoption of language policies projected by the EU and CoE. In this vein, believing the importance of international cooperation for the improvement in 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 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 respect for the students/examinees, fairness, validity, reliability, responsibility and collaboration among the allies involved.

Within the boundaries of '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 ongoing assessment practices.

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 a 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, Holzknecht & 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.

4. The European Standards of Language Testing and Assessment: ILTA

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'.

ILTA bolsters ethical standards in language testing by means of 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 'ough-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.

Besides the Code of Ethics, ILTA also proposes the Guidelines for Practice, whose draft version was firstly introduced at the ILTA meeting held in Ottowa 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, 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.

5. The European Standards of Language Testing and Assessment: ALTE

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. Secondarily, 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.

Bearing these in mind, the current study was employed in order to scrutinize the exploitation of the European guidelines in language testing and assessment practices in non-formal educational settings. Correlatively, the perceived gap in the literature is aimed to be filled with the answer to the following research question:

- 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?

6. Method

The study was conducted by a mixed-methods research design-based investigation of the ongoing testing and assessment practices of English language schools rendering non-formal education in Turkey to the European standards set by the CEFR, EALTA, ALTE and ILTA.

6.1. Participants and setting

40 English language teachers (12 male and 28 female participants who are also test-item developers) from three institutionalized private English language schools offering non-formal education, all of which are the members of ÖZ-KUR-DER (the Association of Private Educational Institutions and Study Center in Turkey) as the most prominent courses renowned for quality in learning and teaching English in Turkey with the highest course attendee capacity and generalizability of the results are recruited. The age range ranks among 18-45 with less than five to more than fourteen years of teaching experience.

6.2. Instruments

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 are listed considering the guidelines proposed

by the EALTA; the Manual recommended by the CEFR; guidelines for practice introduced by the ILTA; and the Code of Practice ascertained by the ALTE. A questionnaire composed of 63 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 63 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 all coded reversely. Then, overall Cronbach's Alpha level for the instrument was evaluated for the context in which the present study conducted was .952.

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. 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.

6.3. Sampling procedures

Following literature review and preparation, the researcher visited the director of ÖZ-KUR-DER to spot the most appropriate English language schools which were also listed

as the members of ÖZ-KUR-DER. Then, the minimum standards for establishing quality profiles in exam were composed, and the pre-selected English language schools were visited one by one.

Quantitative data were gathered from the English language teachers, who were also test-item developers at those institutions, by the 5-point Likert type questionnaire above mentioned in detail. On the other hand, qualitative data were gathered from the directors of each English language schools together with the director of ÖZ-KUR-DER through semi-structured interview sessions. The quantitative data gathered were analyzed by SPSS Version 23.0 whereas the qualitative data were analyzed by constant-comparison analysis, in which researchers were assumed to come up with an emergent fit, albeit not linking data with a pre-determined category (Taber, 2000), and the results were reported below.

7. Results and Discussion

7.1. 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 general paradigm of a sample of leading professionals from selected private institutions were reported in a two-way alternate: (a) the utilization of the European guidelines in language testing and assessment by selected private institutions; and (b) the viewpoints of the directors from those private institutions and ÖZ-KUR-DER on the utilization of the European guidelines in language testing and assessment practices.

7.1.1. The utilization of the European guidelines in language testing and assessment by selected private institutions

In terms of the EALTA guidelines, the overall results 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 predefined 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 also 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 opinions 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 face 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 the 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 the Reference Supplement, though.

Relatively, the results also 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 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.

Regarding the criteria set by the ALTE, 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 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 judgement 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 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) was 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.

In the light of the ILTA guidelines for practice, 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).

To sum up, it was reported by the findings of this study that the selected private institutions did not embrace the European guidelines thoroughly. In terms of the utilization of the European guidelines in language testing and assessment practices by selected private institutions, the private institution B outscored the others in each of the European guidelines. This might be due to the fact that the private institution B was 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.

There is an adoption of the Framework, albeit inefficiently, and the procedures recommended in the Manual and Reference Supplement are not followed although the Framework has a notable influence on language testing and assessment (Coste, 2007). The ELP is to some extent in use as a self-assessment tool in the selected private institutions; however, the ELP is widely implemented around the world via its free access in many languages (CoE, 2011; Little, 2005; Mirici, 2008; Schaerer, 2005).

The lowest mean score was estimated on the utilization of the EALTA Guidelines (M= 3.73). But the exploitation of the EALTA Guidelines to 'frame a validity study' (Alderson, 2010) is recommended. The second lowest mean score was estimated on the utilization of the ALTE Code of Practice (M= 3.74). However, the ALTE Code of Practice is proposed as a cadre for monitoring professional standards in language testing and assessment (Saville, 2005).

It was noted by the findings that another lowest mean score is estimated on the use of standardized tests within selected private institutions (M= 3.45). Herein, 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.

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

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. Secondarily, 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.

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 ongoing 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 skill-based teaching was to be highlighted more, and to be put into use.

With respect to the recommendations for the enhancement of ongoing 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 ongoing testing and assessment practices across the country, the director of 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.

In other respects, for the enhancement of ongoing 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 ongoing testing and assessment practices across the country, the director of A stated that skill-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.

Accordingly, private institutions rendering English language education under the frame of non-formal education are mostly regarded as trading houses merchandizing education. 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 non-formal education of the EU, Bjornavold (2000) suggests that contextual nature of learning, identification of methodological requirements for assessing non-formal learning are to be reconsidered in conducting educational activities on a non-formal basis.

Besides, 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. Additionally, 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).

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. The ratio of auditing is rather low as to that of formal educational settings. Besides, the teachers are asked to fill in the questionnaire rendered by the Ministry of National Education Data Processing Systems (MEBBIS) on an annual basis. However, the results are not sent back to the centers, or even are not announced to the teachers.

8. Conclusions

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 1.500 Turkish Liras per person, 100.000.000 Turkish Liras are spent each year to learn English at private institutions (Karaboğa, 2013). Therefore, there is a need for a validation process for language examinations as the certificates rendered might be invalid for further educational and/or professional purposes.

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. Accordingly, it mushrooms a need for more qualified language teachers together with an increase in the number of in-service teacher training facilities. The language teachers are also expected to become assessment-literate so as to conduct language testing and assessment practices efficiently (Kavaklı & Arslan, 2019).

Setting standards by the adoption of the Framework as the current reality of the ELT professionals, since the Framework is now more than just being 'common' and 'European', albeit internationally recognized worldwide (Mirici & Kavaklı, 2017). Otherwise, the English language teachers are somehow led to continue with the habit of on-going reiteration of the same old story without the reconceptualization of the current EFL curriculum in use. Herein, there is a need for a more practical curriculum enabled through the adoption of the Framework.

Long-term meaningful effects are to be reckoned until acceptable results are achieved in order to ensure 'no tissue rejection' (Holliday, 1992). 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. In this context, there is a need for cooperation amidst the allies, such as private institutions, universities, MoNE, other public and private education centers, agencies and even non-governmental organizations.

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. 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.

Acknowledgements

This research is based on a PhD thesis entitled "CEFR oriented testing and assessment practices in non-formal English language schools in Turkey" submitted to Hacettepe University Graduate School of Social Sciences in 2018.

Endnote

This study was partly presented at GlobELT 2019: 5th International Conference on Teaching and Learning English as an Additional Language, which was held in Kyrenia, Cyprus between the dates of April 11-14, 2019.

References

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal (MLJ), 91*(4), 659–663.

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. (2010). A survey of aviation English tests. *Language Testing, 27*(1), 51-72.

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.

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 2007. [On-line: http://www.coe.int/T/DG4/Linguistic/Source/SourceForum07/D-Coste_Contextualise_EN.doc, Retrieved on 25 September, 2018,]

Council of Europe (CoE). (2000). *Resolution on the European language portfolio*. Adopted at the 20thsession 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). (2005). *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). (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.

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.

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.

Erickson, G., & Figueras, N. (2010). *EALTA guidelines for good practice in language testing and assessment: Large scale dissemination days.* [Online http://www.ealta.eu.org/documents/archive/GGP_dissemination_report.pdf, Retrieved on 15 July, 2016.]

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.]

Holliday, A. (1992). Tissue rejection and informal disorders in ELT projects: collecting the right information. *Applied Linguistics, 13*(4), 403-424.

International Language Testing Association (ILTA). (2007). *Guidelines for practice in English.* [On-line: http://c.ymcdn.com/sites/iltaonline.siteym.com/resource/resmgr/docs/ilta_guidelines.pdf, Retrieved on 20 September 2016.]

Kaftandjieva, F. (2004). *Reference supplement to the preliminary pilot version o the manual fo relating language examinations to the CEF: Section B: Standard setting*. Strasbourg: Language Policy Division.

Karaboğa, K. (2013, 23 December). *İngilizce'ye yılda 100 milyon* [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., & 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.

Kavaklı, N. (2018). *CEFR oriented testing and assessment practices in non-formal English language schools in Turkey*. Unpublished PhD Thesis. Ankara: Hacettepe University.

Kavaklı, N., & Arslan, S. (2019). Towards a continuum from know-how to show-how for developing EFL student-teachers' assessment literacy. *International Online Journal of Education and Teaching (IOJET), 6*(1), 223-232.

Kavaklı, N., & Mirici, İ. H. (2018). The utilization of the European standards for defining educational assessment: teacher-tester attributes and directors' control. *Selçuk University Journal of Faculty of Letters (SEFAD), 40*, 171-190.

Lambert, D., & Lines, D. (2000). *Understanding assessment: Purposes, perceptions, practice.* London, UK: Routledge Falmer.

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.

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.

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.]

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.

Spöttl, C., Kremmel, B., Holzknecht, 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.

Taber, K. S. (2000). Case studies and generalizability: Grounded theory and research in science education. *International Journal of Science Education, 22*, 469-87.

Toncheva, S., Zlateva, D., & John, P. (2017). *Developing as 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. (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/IPOL-CULT_ET(2013)495871_EN.pdf, Retrieved on 17 November, 2016.]

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.

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.
