



## Developing and validating a computerized oral proficiency test of English as a foreign language (Coptefl)

Cemre Isler <sup>1,\*</sup>, Belgin Aydin <sup>2</sup>

<sup>1</sup>Department of Foreign Language Education, English Language Education, Faculty of Education, Fırat University, Elazığ, Turkey

<sup>2</sup>Department of Foreign Language Education, English Language Education, TED University, Ankara, Turkey

### ARTICLE HISTORY

Received: May 19, 2020

Revised: Oct. 13, 2020

Accepted: Jan. 05, 2021

### Keywords:

Language testing,  
Oral proficiency testing,  
Computerized oral proficiency testing,  
English language learners,  
EFL context.

**Abstract:** This study is about the development and validation process of the Computerized Oral Proficiency Test of English as a Foreign Language (COPTFL). The test aims at assessing the speaking proficiency levels of students in Anadolu University School of Foreign Languages (AUSFL). For this purpose, three monologic tasks were developed based on the Global Scale of English (GSE, 2015) level descriptors. After the development of the tasks, it was aimed to develop the COPTFL system and then compare the test scores and test-takers' perspectives on monologic tasks between the COPTFL and the face-to-face speaking test. The findings from these quantitative and qualitative analyses provided substantial support for the validity and reliability of the COPTFL and inform the further refinement of the test tasks.

## 1. INTRODUCTION

Testing students' overall language ability in an efficient manner is one of the primary challenges faced by large-scale preparatory school programs in the universities of Turkey (Aydın et al., 2016). The demands of efficiency often take precedence over in the proficiency tests of these programs and as a result, in most cases, the administrations of oral proficiency tests are not held for reasons of impracticality and difficulty of implementation (Aydın et al., 2016; 2017). That is, the administration of oral proficiency testing is a time consuming and labor-intensive process (Kenyon & Malabonga, 2001; Mousavi, 2007). For example, the employment of a trained interviewer, such as in the face-to-face oral proficiency interviews, brings about its logistical issues when large numbers of test-takers are to be tested. Other practices, such as paired or group testing procedures, also consume much time and attention in the process of the administrations and are most feasible for small-scale assessments (Malabonga, Kenyon & Carpenter, 2005). Thus, the demands for testing speaking make it impractical to systematically measure in foreign language programs. For this reason, many institutions don't even try to test

CONTACT: Cemre Isler ✉ [cemreisler@anadolu.edu.tr](mailto:cemreisler@anadolu.edu.tr) 📍 Fırat University, Department of Foreign Language Education, English Language Education, Faculty of Education, Fırat University, Elazığ, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

speaking skills (Aydın et al., 2016).

Due to not given the same evaluative attention as the other skills, Turkish learners of English do not experiment with the oral language as much as they do with the written language. This situation, in turn, causes a lack of motivation for the achievement of communicative oral skills on the part of the students (Aydın et al., 2016). Harlow & Caminero (1990) articulated this point as: “If we pay lip service to the importance of oral performance, then we must evaluate that oral proficiency in some visible way” (p.489). Indeed, most English language instructors in Turkey are well aware of the importance and necessity to test directly the speaking skill in the proficiency tests. Teachers, however, are confronted with the fact that there does not exist an oral proficiency instrument or a model that is easy to implement for a large group of students in terms of time and logistics. One of the current studies on this topic was conducted by Aydın et al. (2016) in which they carried out a series of interviews with the administration of twelve schools of foreign languages in Turkey. There were two purposes that leading this study. First, it was aimed to explore the practices used by the universities to prepare reliable and valid language proficiency and to discuss the feasibility of these practices in their contexts. Second, it was aimed to collect opinions from these state universities in Turkey about the use of computer-assisted assessment techniques in the assessment of language proficiency, as well as to identify the existing practices if there any. The findings of the study revealed a detailed picture of the present practices of universities concerning language proficiency tests. The most prominent findings of the study showed that (1) all institutions believe the importance of including four skills in a proficiency test; namely reading, listening, writing, and speaking. Yet, most of them cannot test the speaking skill due to practical reasons; (2) most of the institutions refer to not having sufficient human resources and technical equipment for the preparation, administration, and assessment procedures of proficiency tests. These tests are mostly prepared and administered by the instructors assigned for this job or volunteers to do it. The number of staff in testing units who received education in assessment and evaluation is quite low; (3) they also state experiencing certain problems in the administration of proficiency tests. Accordingly, it is not possible to pilot the tests due to time limitations both for administration and assessment procedures. Due to the high number of students, tests are provided in multiple-choice format and the statistical analyses of test results are not done by experts in most institutions because of the reasons mentioned above. Within the purpose of this study, particularly about the speaking skill, the data gathered from the leading universities of Turkey clearly show that among all skills, testing oral proficiency is referred to as the most problematic one which results in not testing at all. The results, all in all, clearly depict the lack of agreed content and the administration and assessment of the framework for proficiency tests. However, establishing certain standards in foreign language education seems inevitable to catch up with the developed countries with regard to internationally recognized language tests in terms of validity, reliability, and usability. In this regard, all the universities that participated in the study emphasize the necessity of establishing certain standards in foreign language education. Also, all of them except one state that they support the idea of developing a nation-wide proficiency test by using technology.

When we have a look at the studies on educational technology, we see that with the recent advancements in computer technology, the use of computers in the delivery of oral proficiency tests has begun appealing due to its potential benefits such as increased reliability of the test as a consequence of the standardization of test delivery process, more efficient test administration and the flexibility in the delivery of tasks (Mousavi, 2007; Zhou, 2015). Although recent advances in computer technology have promoted the computer delivery of oral proficiency tests, the absence of an interviewer has resulted in concerns about the validity of using them as a replacement for face-to-face speaking tests (Zhou, 2015). Accordingly, the most ubiquitous concern was that test-takers’ performance on a computer-based speaking test may not reflect

their ability measured by face-to-face speaking tests in which test-takers are required to interact with an interviewer (Zhou, 2008; 2015). Examining this issue of importance, since it concerns fundamental questions of test validation, i.e. to ensure the score interpretations (Zhou, 2008). So, there has been a call for more research on comparing computer-based tests with conventional face-to-face speaking tests. Given that the score equivalence is significant and should be established prior to the interpretations of computer-based speaking tests, in the present study it was attempted firstly to develop the Computerized Oral Proficiency Test of English as a Foreign Language (COPTTEFL) and then investigate the equivalence of the semi-direct (COPTTEFL) and the direct (face-to-face) versions of a test of oral proficiency. The present study is, therefore, comparability research and it primarily relied on concurrent validation which focuses on the equivalence between test scores. However, this study argues that examining the relationship between test scores only through concurrent validation might provide insufficient evidence as to whether the COPTTEFL measures what it intended to measure. It suggests demonstrating the validity of the test from multiple perspectives. In this respect, it suggests that test-taker attitudes might represent an important source to obtain a deeper understanding with regard to the construct validity and face validity of the tests. If test-takers' attitudes towards the test seriously affect their scores, the scores may not reflect their real language ability, which the test is intended to assess and consequently, the test would lack construct and face validity. With these purposes, it was firstly aimed to develop a computer-based speaking test system, namely the COPTTEFL which would be established on a framework of test validation.

### **1.1. A framework for validating a speaking test**

The most useful starting point for the test development is to have a framework of validation to support the claims made for the tests. If the study is to establish whether the test is valid as a testing instrument, it is essential to utilize a framework of validation in order to collect data systematically and objectively. The socio-cognitive framework (Weir, 2005) for validating the test was used in the present research. It was operationalized from the initial stages in the development of the test of speaking to the comparability of scores by each mode of testing. Several frameworks for language test validation have been proposed by earlier theorists, but as put forward by O'Sullivan (2011a), they have been unable to offer an operational specification for test validation. The approach taken by Weir (2005), however, defined each aspect of validity with sufficient detail as to make the model operationalizable for each of the four skills (O'Sullivan, 2011b).

The socio-cognitive framework for validating the speaking test (Weir, 2005) was used as the major reference by which the speaking tasks of the study were developed. The framework offers a guideline for validating the speaking tests by demonstrating the steps that need to be followed for validity and reliability concerns. The essential components to be investigated in the framework are as follows: (1) Test-taker characteristics, (2) Theory-based validity, (3) Context validity, (4) Scoring validity, (5) Criterion-based validity, (6) Consequential validity.

Firstly, test-taker concern has been raised by Weir (2005) and it was argued that it is directly related to the theory-based validity since test-taker characteristics have an impact on the way test-takers process the test task. He stated that physical, psychological, and experiential differences of the individuals should be considered during the test development process so that bias for or against a particular group can be avoided. Secondly, theory-based validity is related to considerations regarding how well a test task correlates with cognitive (internal mental) processes resembling those which language users employ when undertaking similar tasks in non-test conditions. Thirdly, context validity is related to the appropriacy of the contextual properties of the test tasks to assess specific language ability. Moreover, scoring validity is concerned with the extent to which test results are consistent with respect to the content

sampling and free from bias. Criterion-based validity is about the relationship between test scores and other external measurements that assess the same ability. Finally, consequential validity refers to the impact of tests and test scores interpretations on teaching, learning, individuals, and society.

The present study only focused on theory-based validity, context validity, and scoring validity. The other aspects of the framework; test-taker characteristics, criterion-based and consequential validity were not investigated. This was decided on for the reason that it was beyond the scope of the study to collect data on all components due to time constraints. Therefore, only those included in the study were discussed in the following part of the study.

## 1.2. Theory-based validity

Theory-based validity, construct validity, or later renamed as cognitive validity (Khalifa & Weir, 2009), is one of the components of Weir's (2005) socio-cognitive framework for validating language tests and concerned with the internal mental processes. In relation to the cognitive processes elicited from test-takers, Field (2013) argues that the main concern is not whether the tasks are close to an actual speaking or listening event, but whether these tasks require test-takers to employ the internal mental processes that a language user normally undertakes in similar tasks during non-test conditions. Reflecting on the representatives of the mental processes in test tasks is the main concern for cognitive validity. Therefore, the focus in studies of cognitive validity is not on the speech produced by the test-taker, but rather the mental processes that a test-taker undertakes in speech production during a speaking test. At this point, the relationship between theory-based validity and context validity is a symbiotic one. The context in which the test task is presented has an impact on the mental processes of the test-taker. For example, the mode of input, whether it is listening to the dialogue or looking at pictures will influence how the test-taker conceptualizes and processes these messages as pre-verbal messages (Zainal Abidin, 2006). The speaking skill descriptors provided by the Global Scale of English (GSE, 2015) were used in the present study to define the language construct and determine the target sub-skills of the construct.

### 1.2.1. GSE descriptors for the speaking skill

After Messick's (1989) challenge against the traditional view of validation, validity is not seen as a characteristic of a test, but a feature of the inferences made on the basis of test scores. The focus here is the test score or the results of the test since this is what is used to make interpretations about test-takers' ability (Chapelle, 2013). As stated in Chapelle (2013), in current approaches, scores are interpreted with regard to pre-determined standards of knowledge. For example, the increasingly used Common European Framework of Reference for languages (CEFR, Council of Europe, 2001) represents an ordered set of statements through six common reference levels (A1, A2, B1, B2, C1, C2; *ranging from lowest to highest*) that describing language proficiency. It is claimed, for example, that a speaker assessed as meeting the standard for level B1:

“Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst traveling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes, and ambitions and briefly give reasons and explanations for opinions and plans” (Council of Europe, 2001, p.24),

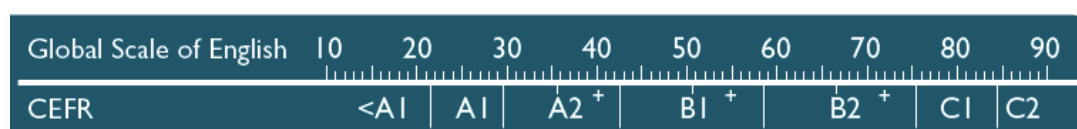
while a speaker at B2:

“Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint

on a topical issue giving the advantages and disadvantages of various options” (Council of Europe, 2001, p.24).

The development of a test instrument begins with such a set of standards (Chapelle, 2013). These may be rather general, as in the case of the CEFR, or more granular, as in the GSE. The GSE proficiency scale was created with reference to the CEFR, but the main difference between the GSE proficiency scale and the CEFR proficiency scale stems from its granular structure (see Figure 1).

**Figure 1.** Global Scale of English aligned with the CEFR.



As shown in Figure 1, the GSE presents a more granular measurement of proficiency within a single CEFR level (GSE, 2015). It is a proficiency scale from 10 to 90 and defines what a learner can do across four skills at a specific GSE range. For example, a language learner at GSE range 27 “can understand a phone number from a recorded message, but a learner at 74 “can follow an animated conversation between two fluent speakers” in listening skill. As for reading skills, a learner at 43 on the scale “can understand simple technical information (e.g. instructions for everyday equipment)” whereas a learner at 58 “can recognize the writer’s point of view in a structured text”. As for speaking skills, a learner at 42 “can give a short basic description of events and activities” while the ones at 61 “can engage in extended conversation in a clearly participatory fashion on most general topics”.

Most of the preparatory programs in Turkey use the CEFR as a proficiency scale where the learner proficiency is classified from A1 (low basic) to C2 (fully proficient) (Council of Europe, 2001). However, in the 2014-2015 academic years, Anadolu University School of Foreign Languages (AUSFL) moved away from CEFR towards the GSE which is psychometrically aligned to CEFR (GSE, 2015). The reason for this shift from CEFR to GSE was explained as:

“The wide proficiency ranges covered by each of the 6 CEFR levels (from A1 to C2) made it difficult for everybody to agree on the exact nature of each proficiency level. Considering the nature and difficulties of the language learning process, especially in a foreign language context, the inability to demonstrate how much progress has been achieved and how much more remains might be a demotivating factor. The time it takes for students to move up from one level to another varies greatly depending on their starting level, the amount of exposure to the language, their context, mother tongue, age, abilities and a range of other factors. For this reason, it is difficult to estimate how much time is needed to pass from one CEFR level to the next, especially in a context where input is mainly limited with the classroom boundaries. These limitations, in addition to the lack of clarity on how to interpret the CEFR levels, required searching for a different proficiency framework which resulted in the discovery of the Global Scale of English (GSE), a psychometric tool” (Aydin et al., 2017, p. 308-309).

The curriculum of the speaking course was designed based on the GSE (2015) Learning Objectives between 51-66 levels. 66 on the GSE proficiency scale, which corresponds to the initial stages of B2 in the CEFR was established as the optimum point to be reached by the end of the program. The reason why 66 was determined as an exit level was that “considering the entry-level and the length of time available for both in and out-of-class study, 66 was determined to be an achievable point on the GSE” (Aydin et al., 2017, p. 311).

Since the program aims to give general English from 51 to 66 on the GSE scale, it was therefore decided to take the range between 51-66 levels for the speaking skills as a basis for the test tasks developed in the present study. The fact that the GSE (2015) identifies language



proficiency in different levels and offers illustrative descriptors of “can-do” activities at each range of a proficiency level makes it a useful reference in task design especially when a specific range is targeted for a task. These descriptors are used in the study to guide the alignment of the tasks to the different proficiency levels.

### 1.3. Context Validity

Context validity, which is often named as content validity (i.e. Fulcher & Davidson, 2007) is related to the context coverage, relevance and representativeness. The contextual components of the test tasks in the study are examined based on the aspects of context validity for speaking proposed by Weir (2005). According to Weir (2005), it is “the extent to which the choice of tasks in a test is representative of the larger universe of tasks of which the test is assumed to be a sample” (p.19). This description implies that task characteristics and settings of tasks should reflect “performance conditions of the real-life context” as much as possible (Shaw & Weir, 2007, p.63).

Weir (2005) notes that in test development, various elements regarding the task and administration setting, as well as task demands in terms of linguistic characteristics and speakers should be taken into account to develop a theoretically sound basis for the choices made with respect to contextual features of the test tasks. Therefore, presenting as much evidence as possible for each of these elements will provide test developers with pieces of evidence to validate the choices they would like to make about test-takers based on their test performances. In this sense, the tasks developed for the purposes of the current study were ordered according to their assumed difficulty based on the GSE scale descriptors and targeted specific range in the scale (between 51-66 levels, see Section 2.3.2 for further information). The GSE scale descriptors provide an opportunity for producing a wide range of speech functions as describing, comparing, elaborating and expressing preferences, explaining, and justifying opinions. The current test taps into these various functions since different functions require different kinds of cognitive processing and may increase/decrease task difficulty (Galaczi & ffrench, 2011).

### 1.4. Scoring validity

Scoring validity is concerned with all test aspects that can influence scores’ reliability. Zainal Abidin (2006) highlights that scoring validity is an inevitable aspect of test validation procedure since the scores obtained from the tests may not be totally due to their performances, but influenced by other factors i.e. sources of error. Such problems of inconsistency can threaten the validity of the test and lead to the involvement of construct-irrelevant variance in the testing process. Therefore, it is identification and minimization of such errors of measurement that test developers should concern for the reliability of the scores produced by a test.

In testing speaking, rating is an important factor affecting the reliability of the test. It includes criteria/rating scale, rating procedure, raters, and grading and awarding. The chief concern in the testing of speaking, in this sense, is rater reliability and how scores are awarded based on a rating scale (Zainal Abidin, 2006). As for the investigation of the scoring validity in the study, it was examined how well the COPTEFL scores and the face-to-face speaking test scores are compared in terms of inter-rater reliability, order-effect and test scores.

### 1.5. The present study

When we review the research on the use of computers in oral proficiency testing, it is seen that the studies have focused largely on correlations or analyses of test outcomes/products including test scores (Jeong, 2003; Kiddle & Kormos, 2011; Öztekin, 2011; Thompson, Cox & Knapp, 2016) underlying constructs of test-taker language output in different modes of tests (Zhou, 2015), and test-taker reactions (i.e. attitudes) (Joo, 2008; Kenyon & Malabonga, 2001; Qian,

2009). As O’Loughlin (1997) states while these approaches have offered valuable insights into the comparability issue, there is a need to complement them with other perspectives as well. In particular, apart from investigating test outcomes or products, limited attention has been paid to “the examination of test design, test taking and rating processes and how an understanding of these components of assessment procedures may provide the basis for a more complex comparison between the two kinds of tests when combined with the analysis of test products” (p. 72). The perspective that O’Loughlin (1997) addressed above is the methodological approach taken in the current study. Particularly, apart from investigating comparability of test scores and test-taker attitudes obtained from open-ended questions, the study placed emphasis on the examination of the test development process including test design, development and administration stages. To date, such an approach has been seldom adopted in comparability research. Notable examples of studies combining a focus on process and product in testing speaking are O’Loughlin (1997) and Mousavi (2007). This study differs from O’Loughlin (1997) because its aim was not to compare a tape-based speaking test with a face-to-face speaking test. It also differs from Mousavi (2007) because its aim was not to compare a computer-based oral proficiency test with a face-to-face speaking test (International English Language Testing System, IELTS) that already was in use. Instead in the present study, first, a computer-based oral proficiency test format was developed and then a face-to-face version of the test was created to provide contrast and the test-method effect. By attending to both process and product, it was aimed to offer greater insight into construct validity and thus establish a stronger basis from which to compare test scores and attitudes towards both test delivery modes. The research questions that guided the present study are as follows:

1. How well are the COPTEFL scores and the face-to-face speaking test scores compared in terms of (a) inter-rater reliability; (b) order-effect; (c) test scores?
2. What are the attitudes of test-takers in relation to the test delivery modes?

## **2. METHOD**

### **2.1. Research context and participants**

The study was conducted at Anadolu University School of Foreign Languages (AUSFL), Turkey. The school is a preparatory program that aims to equip the students with the necessary language skills in order to follow the academic education in their departments. The curriculum of the program is designed to help students to be able to reach that exit proficiency level required to be accepted as successful and constructed based on the GSE Learning Objectives (2015) between 51-66 levels. This test was designed for newly arrived students to AUSFL who have finished their high school or the students who are studying in preparatory schools of the university and have to take an exit exam to demonstrate that they have gained sufficient proficiency in English for academic study in their departments. The participants of the study were forty-five non-native speakers of English whose first language is Turkish. Participation in the study was on a voluntary basis.

### **2.2. Test development team**

#### **2.2.1. Item writers and raters**

There were eight-item writers in the study. Item writers were also the raters. They are professional instructors who work full-time for the testing institution in AUSFL. They are experienced teachers of similar students and they have relevant experience for teaching speaking, writing speaking tasks for proficiency exams and assessing students’ speaking proficiency.

#### **2.2.2. Editors**

In the editing committee, there were four experts who were asked for opinions about the written

items. One of the experts was a professional item writer and rater who did not participate in producing items and scoring. Another one was experienced in teaching speaking and language testing in Anadolu University Foreign Languages Department. The others were subject experts, the researcher was one of them.

### 2.2.2. Software developers

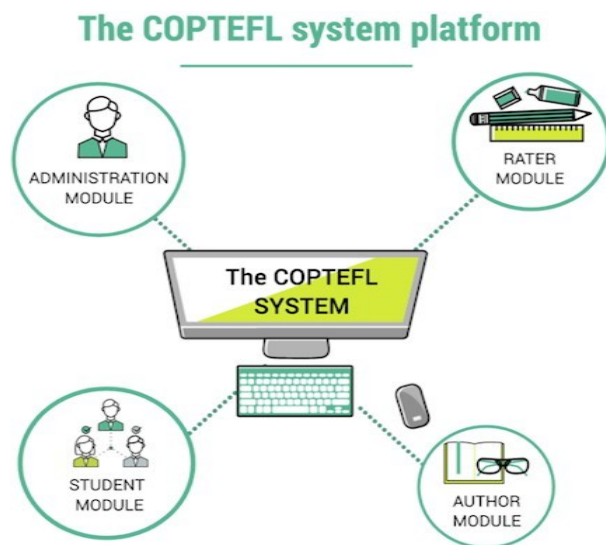
Two software developers, who have the necessary formal professional qualifications, developed the COPTEFL system. One of them worked for the programming of the system and the other worked for the web-page design.

## 2.3. Instruments

### 2.3.1. The COPTEFL system platform

The COPTEFL is a computer-based speaking test of general proficiency designed for adult learners of English as a Foreign Language (EFL). It uses the Web as its delivery medium. It was designed to offer users an alternative to face-to-face oral proficiency tests. The COPTEFL system platform comprises four types of users as (1) Administrator, (2) Author, (3) Rater, and (4) Student. Each user has its own module and is only allowed to access the information and functions they are given permission to access. In order to keep the system reliable, only the administrator(s) have access to other modules (see [Figure 2](#)).

**Figure 2.** *The COPTEFL system platform.*



In order to keep the system reliable, only the administrator(s) have access to other modules. Each module and its functions were explained below:

#### 1. Administration Module:

This module is used by the administrator(s) in order to manage the system platform. Accessing this module allows direct access to all components of the COPTEFL system. The functionalities built into this module include:

- (a) Managing users (authors, raters and students) i.e. allowing or restricting user access, accessing the data of a user in particular, editing users' personal data, setting user deadlines (see [Figure 3](#)).



Figure 3. Administration module: Managing users.

COPT EFL  
New Generation Oral Proficiency Test

Home

Students

Instructors

Test

Questions

Settings

Please fill in the textbox

Photo  
Choose File No file chosen

Mail:

Phone:

Type:  
Instructor

Save

Instructor List  
Add Instructor  
Send Message  
All Messages

(b) Setting the time for the test and its announcement to the student module for the test-taker registration (see [Figure 4](#)).

Figure 4. Administration module: Setting time for the test.

COPT EFL  
New Generation Oral Proficiency Test

Home

Students

Instructors

Test

Questions

Settings

Create Test

Name:

Date:

Time:  
18:00

Student Count:



Last Registration Date:

Location:

Create Test

- (c) Allowing test to start on the scheduled time,
- (d) Monitoring the processes of test item creation and rating, and sending messages to the authors/raters,
- (e) Control over the written test items i.e. editing or deleting before they are included in the item pool by the system automatically (see [Figure 5](#)).

**Figure 5.** Administration module: Control panel for the written test items.

Part 1   Part 2   Part 3						
Number	Image	Label	Question	Writer	Proofreader	Proofreader Review
1		Daily life	Describe this picture and explain why it is important to have a good night's sleep.	Abdulkadir Durmuş	Proof Reader	Approved
2		Education - school life	Describe this picture and explain why it is important to have a university education.	Abdulkadir Durmuş	Proof Reader	Approved

- (f) Monitoring the test questions for each student before and after the administration of the test,  
 (g) Changing the test questions before the test starts,  
 (h) Accessing the students' answers to the questions during the test administration,  
 (i) Accessing the test results (see [Figure 6](#)).

**Figure 6.** Administration module: Accessing to the test results.

17 Mayıs 5/17/2018	Time: Location: Last Registration Date:	14:00 c 107 5/17/2018				
All						
E-mail	Part Number	Rater Grade	Inter Rater Grade	Proofreader Grade	Result Grade	Result
hasanalibuyuksahin@hotmail.com	Part 1	65	70	0	62	Passed
	Part 2	70	55	0		
	Part 3	55	55	0		
tekdemir1558@gmail.com	Part 1	55	85	75	60	Failed
	Part 2	55	85	70		
	Part 3	55	70	50		

## 2. Author Module:

This is the module used by the item writers to develop and edit tasks for the test item pool (see [Figure 7](#)).

Figure 7. Author module: Adding questions.

The screenshot shows the 'New question' form in the COPT EFL author module. The form is titled 'New question' and is located on a page with a dark blue sidebar on the left. The sidebar contains navigation links: Home, Test Bank, Rating, and Review. The main content area has a teal header with the text 'New question'. Below the header, there are several input fields and buttons. The 'Select Part' field is a dropdown menu with 'Please select...' as the placeholder. The 'Select Label' field is also a dropdown menu with 'Please select...' as the placeholder. The 'Photo' field has a 'Choose File' button, a '(500x500)' label, and an 'Upload' button. The 'Question' field is a large text area. The 'Note' field is a smaller text area with '(If any)' below it. A 'Save' button is located at the bottom of the form.

The tasks written by the item writers are shown in the administrator’s module so that any necessary final changes can be made before the tasks become ready for the test. Item writers can only manage their own tasks and cannot access the tasks developed by other writers. However, they can see the total number of tasks for each part in the item pool.

### 3. Rater Module:

This is the module through which raters can monitor the students’ tests to score. These tests are assigned to raters by the computer automatically. Each rater is also an inter-rater. Raters do not know for which test they are assigned as rater or inter-rater. They just give the scores for each test that showed up on their module. Only administrator(s) can access the data of a person about for which test s/he was a rater or inter-rater. Scoring is anonymous on the system. The identities of the students remain confidential. Each task is delivered successively to score. Depending upon the extent of the discrepancy between scores, two or three rater scores were compared to get more accurate results. In AUSFL’s rating system if the scores by two raters are discrepant by more than ten points, a third rater independently scores. The score of one of the two raters whose score is close to the third rater is accepted as valid. This procedure is adapted in the ratings of the study. A proofreader who is the administrator gives scores as a third rater in order to reach a consensus in the ratings among two raters. The scores of the one whose scores are close to the proofreader’s are accepted as valid by the computer and therefore, the final score comes from the average score of these two raters.

In the development of the rating scale, an existing scale -which was developed by a formal testing body in AUSFL was adopted. But, it was modified since it involved “interaction” as a rating element. Because the tasks were monologic ones, “interaction” would be useless. The process in specifying the procedures for scoring started with expert judgments and evaluations. Appropriate changes were made based on those decisions by the experts and therefore, the tasks are decided to be scored according to the following assessment criteria: (1) Pronunciation, (2) Fluency, (3) Grammar range and accuracy, (4) Adequacy of vocabulary for purpose and (5) Task fulfillment. In the rater module, each criterion is shown with its explanation on the page

and raters see the criteria section while they are listening to the answers. They can give scores at the same time they are listening to (see Figure 8).

Figure 8. Rater module: Giving scores.

The screenshot shows the 'Please Listen to Answer' and 'Please Rate' sections of the Rater module. The 'Please Listen to Answer' section includes an image of four people socializing, a question to describe the image and explain its importance, and a 'Current point = 65' indicator. The 'Please Rate' section shows rating criteria for 'Task fulfillment' and 'Fluency'.

**Task fulfillment :**

- 4 - EXCELLENT --- Relevant and adequate answer to the task set.
- 3 - GOOD --- For the most part answers the task set, though there may be some gaps or redundant information.
- 2 - FAIR --- Answer of limited relevance to the task set. Possibly major gaps in treatment of topic and/or pointless repetition.
- 1 - POOR --- The answer bears almost no relation to the task set. Totally inadequate answer.

**Fluency :**

- 4 - EXCELLENT --- Speech is effortless with speed that approaches that of a native speaker.
- 3 - GOOD --- Speaks smoothly with little hesitation.
- 2 - FAIR --- Speech is slow and often hesitant and jerky. Sentences may be left uncompleted, but speaker is able to continue, however haltingly.

When they complete marking, the computer shows the total grade that a student gets after the scoring process and then the rater can submit the score (see Figure 9).

Figure 9. Rater module: Completing marking.

The screenshot shows a table of completed marking tasks. The table has columns for Image, Part Number, Question, Status, Point, and Rating. Below the table, there is a 'Current point = 65' indicator and a 'SUBMIT' button.

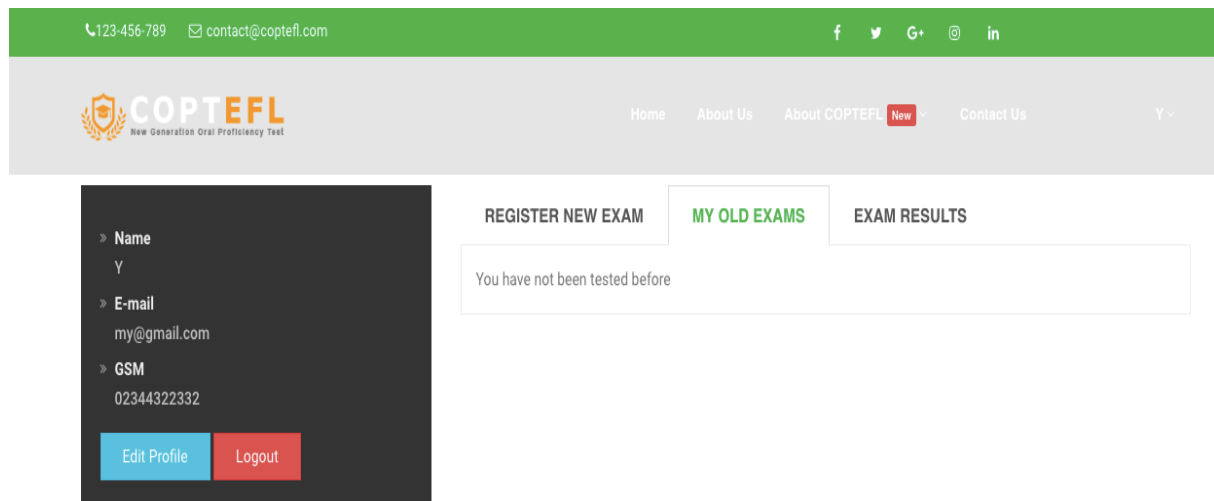
Image	Part Number	Question	Status	Point	Rating
	1	Describe this picture and explain why it is important to have a good night's sleep.	Rated	60	↩
	2	Look at these two pictures and explain whether the fathers or the mothers should spend more time with their children.	Rated	65	↩
	3	1. Talk about an unforgettable moment in your life. 2. How did it affect your life? 3. Do unforgettable moments shape your future life?	Rated	70	↩

Current point = 65 SUBMIT

#### 4. Student Module:

This module has two functions: (1) to deliver the test and (2) to publish the test scores achieved by the students. The students who registered to the system log into the delivery platform in order to start their test. When raters and inter-raters submit their scores, the computer averages the grades and publishes them on students' own page. The students can log into their accounts and see the grades they get for each criterion and their final grade on this module (see [Figure 10](#)).

**Figure 10.** *Student module.*



#### 2.3.2. Speaking tasks

The speaking tasks used in the present study were created by the item writers. Three speaking tasks were developed in order to evaluate students' general proficiency with regard to oral competency. Since the test was intended to be used as a general proficiency test, the tasks were prepared according to the Global Scale of English (GSE) 'can do' statements from 51 to 66 levels (see some examples from the range 51-66 below):

- 51 Expressing and responding to feelings (e.g. surprise, happiness, interest, indifference).
- 53 Comparing and contrast alternatives about what to do, where to go, etc.
- 60 Justifying a viewpoint on a topical issue by discussing pros and cons of various options.
- 62 Constructing a chain of reasoned argument.
- 66 Developing a clear argument with supporting subsidiary points and relevant examples.

The COPTEFL included monologic tasks that can elicit individual discourses without the test-takers' interacting with an interlocutor. These tasks were discourse type tasks. The first task was a description and giving an opinion task. In this task, students were required to describe the picture and then give/express/justify an opinion related to the picture. The second task was a comparison task. In this task, students were asked to look at two pictures and choose one and provide a reason for their choice. The final task was a discussion task in which students were required to justify a viewpoint.

The final version of the task types was based on the pilot test. In this phase, the prototype tasks were piloted in a face-to-face test with small groups of learners in order to find out which tasks do not work as planned, and which should be included or excluded after revision. Announcements explaining that students had the chance of testing their speaking skills were made at the school. Six volunteer AUSFL students participated in the study. The researcher conducted the test and rated for the scores. The analysis showed that some of the questions might elicit a small range of language and repeated answers from the students. The presence of



general questions such as “Why is it important to celebrate national holidays?” provided plenty of scope for answers. This, in turn, caused the detailed question related to the general one such as “Talk about a national holiday celebrated in your country” to be covered in advance. Therefore, the necessary changes and modifications were made after discussing problematic items with the team of test writers. According to this revision, test tasks were redesigned. Item writers and editors were asked for their opinions with regard to the final version of the task types. After getting their approval, test tasks were written. The editors edited the written tasks and the final version of the tasks was entered into the COPTEFL system. After this process, the COPTEFL system was ready to test.

### 2.3.3. Face-to-face speaking test

The tasks and the instructions used in the face-to-face speaking test were the same as the COPTEFL (for an example task for the face-to-face speaking test see [Figure 11](#)).

**Figure 11.** An example for Task 2.

**PART 2:** I will ask you to compare two pictures, choose one and provide reasons for your choice. You will have 15 seconds to think about your answer and 90 seconds to reply.

Look at these two pictures and talk about which of these travelling ways you would prefer to choose.



### 2.3.4. Open-ended questions

Open-ended questions were designed according to the opinions of the two subject experts. Since the aim of the study was to find out the usability of the COPTEFL in comparison with a face to face equivalent, the questions were targeted to depict attitudes of students towards the COPTEFL system and its perceived advantages and disadvantages with regard to a face to face speaking test. Therefore, open-ended questions were used for investigating test-taker attitudes towards testing speaking in the COPTEFL and the face-to-face mode. Test-takers were first asked to evaluate the COPTEFL system and then, state their preferences by making comparisons between the two modes. The questions were:

1. How do you evaluate the COPTEFL system as a speaking test delivery medium?
2. What are the advantages and disadvantages of the COPTEFL when compared to the face-to-face speaking test?
3. What are the advantages and disadvantages of the face-to-face speaking test when compared to the COPTEFL?

## 2.4. Data collection procedure

### 2.4.1. A priori construct validation: Processes followed in the development of the tests

In an effort to adapt the best practices in language test development in the present study, Bachman & Palmer’s (1996) stages in test development were followed with slight modifications in order to more suitably fit the purposes of the research. In their book, test

development was organized into three stages as design, operationalization and administration.

### 1. Test design and operationalization

(a) Developing test task specifications: The purpose for which the test would be used and the target population for the test were our starting point in designing test specifications. Having determined this, relevant literature was reviewed in order to find out what language would be needed by test candidates in the case of oral proficiency tests. From this consultation, similar tasks and texts were sampled to arrive at a manageable test design. Then, with the help of the eminent experts in the field, draft specifications and sample tasks within which the test might be constructed were designed. Since the principal user is probably the test writers, the team of test writers was then asked for their opinions about whether the draft specifications and sample tasks were appropriate for the purposes of the test and the target population. They revised the tasks and specifications, and discussed whether the tasks would work, that is whether each task which was intended to assess a particular aim actually would do so. Many of the responses to whether tasks would work or not gave the impression that a trial on a small group of learners who are similar in background and language level to the target population would provide helpful insights in understanding the kind of language being elicited for each task.

*Trialing for test tasks:* The development of the speaking tasks process consisted of various stages such as the selection of task types, writing of task items, consulting with experts, and pilot tests. The final version of the task types was based on the pilot test. In this phase, the prototype tasks were piloted in a face-to-face test with small groups of learners in order to find out which tasks do not work as planned, and which should be included or excluded after revision. Announcements explaining that students had the chance of testing their speaking skills were made at the school. Six volunteer preparatory program students participated in the study. The researcher conducted the test and rated for the scores. The analysis showed that some of the questions might elicit a small range of language and repeated answers from the students. The presence of general questions such as “Why is it important to celebrate national holidays?” provided plenty of scope for answers. This, in turn, caused the detailed question related to the general one such as “Talk about a national holiday celebrated in your country” to be covered in advance. The result was that, however thoughtfully designed to avoid pitfalls, some of the questions failed to elicit the targeted responses. Therefore, the necessary changes and modifications were made after discussing problematic items with the team of test writers. According to this revision, test task specifications were redesigned to generate test tasks.

Once a coherent system was created for specifications and tasks whose parts fitted together, item writers and editors were asked for their opinions with regard to the final version of the task types. After getting their approval, test tasks were written. The editors edited the written tasks and the final version of the tasks was entered into the COPTEFL system. After this process, the COPTEFL system was ready to test.

(b) Writing test tasks and instructions: After making explicit any constraints in test design, test writers began writing tasks and instructions with the test’s specifications. The writers needed to find suitable communication activities for the tasks such as expressing an opinion on an issue, a view by contrasting it with other possible views, or discussing ideas. The writers also needed to find pictures that serve the purpose of the task. After completing the test writing process, each writer made responsible for editing another writer’s set of tasks. Once their editing process concluded, tasks became subject to a number of reviews before they reached their final draft stage. Two editors revised the items and assembled them into a draft test paper for the consideration of other editors. These editors examined each item for the degree of match with the test task specifications, ambiguities in the wording of the items, and match between the questions and pictures. The changes made after editing processes were reported and shared with the team of test writers.

(c) Specifying the procedures for scoring: In this process, considerable effort was put into developing a practical analytic scale for decision making in which there is less to read and remember than in a complicated descriptor with many criteria and unfamiliar technical terminology. Once the scale was modified, it was then refined by raters who use it so that they understand the meaning of the levels with regard to each particular feature in the scale.

(d) Software development: In developing testing software, there are some key requirements of the standard steps taken by the researchers (Mousavi, 2007; Shneiderman, 2004; Zak, 2001) as (1) Analysing (defining the problem), (2) Choosing the interface, (3) Coding, (4) Test and debug, and (5) Completing the documentation. In this study, the same standard steps were followed. These steps were explained briefly below.

STEP 1: Analysing (defining the problem): This pertains to the definition of a problem in any research project. In this step, a statement of the problem was presented to provide guidance to the rest of the programming steps. That is to say, what exactly the programmer wants to achieve with this programming was stated in this step. The core problem that drew this study was the low degree of the practicality of administering oral proficiency tests to a large group of preparatory program students through the use of a live face-to-face interview. In this step, meetings with the administrator and instructors were held in order to determine the requirements of the COPTEFL system i.e. who would be the rater, whether raters would write items, what to include item writing and rating pages in the system. These were general questions that were answered during the analyzing phase. In order to translate requirements into design, meetings between the researcher and the software developer were held twice a week. They analyzed the requirements of the system for the possibility of incorporating to the COPTEFL system program.

STEP 2: Choosing the interface: The interface of any computerized test involves the actual objects the test-takers see and deal with during a testing session. These objects may consist of videos, text boxes, command buttons, animations, progress bars, date/time indicators, and so on. Here, the key to developing a good user interface is to have a complete understanding of the target user (Luther, 1992). As suggested in Mousavi (2007), this understanding may be achieved by a process referred to as user task analysis where the developer assumes himself/herself as the target user and identifies a series of possible scenarios to come up with the most convenient one. The relationship between the application interface and test-takers is important because, as Fulcher (2003) states interface design can be the threat of interface-related construct irrelevant variance in test scores, and therefore should be avoided. For this purpose, he identifies a principled approach in the development of a good interface design. This approach includes three phases as (1) planning and initial design, (2) usability testing, and (3) field testing and fine-tuning. The present study followed these phases in choosing the interface. Each was explained below.

Phase 1: Planning and initial design. This involved hardware and software considerations, navigation options, page layout, terminology, text, color, toolbars and controls, icons, and the rest of the visible objects on a typical computerized test.

Phase 2: Usability testing. This included activities such as searching for problems and solutions, selecting test-takers for usability studies, item writing and banking, pre-testing, try-out for scoring rubrics.

Phase 3: Field testing and fine-tuning. This consisted of try-out for the interface with a group of samples drawn from the target test-taking population and also making sure that the logistics of data collection, submission, scoring, distribution and retrieval, and feedback would work as planned. This phase provided an opportunity to trial and test for (possible) variation in the appearance of the interface across sites, machines, platforms, and operating systems.

**STEP 3: Coding:** Coding is the translation of the algorithm into a programming language. It is generally the most complex and time-consuming step in the development of the computerized tests. Once the planning of the application and the building of the user interface were complete, programming instructions were written to direct the objects in the interface on how to respond to events. After deciding on the system design requirements, the COPTEFL system was divided into four modules as (a) administration module, (b) author module, (c) rater module, and (d) student module. Coding was developed according to the modules. The code was developed based on the needs of the program from scratch, and this stage was the most challenging part and took the longest time.

**STEP 4: Test and debug:** Debugging is the process of tracking down and removing any errors in the computer program. Errors in a computer program could be the result of typing mistakes, flaws in the algorithms, or incorrect use of the computer language rules. Testing and debugging step was an inevitable part of the operation because of the complexity in the coding phase of the programming as well as the possible persistence of syntactic anomalies in the programming language. Caution must be exercised at this stage for possible problems. After the code was developed, the system went through a pilot study to see if it was functioning properly. The researcher and the developer assessed the software for errors and document bugs if there were any. The developer did the necessary changes to the system due to the results.

**STEP 5: Completing the documentation (or distributing the application):** This step included developing an installable setup file along with all its components for new users and new platforms. That is, all the materials that described the program were compiled to allow other people, involving test users, raters, administrators, and item writers to understand the scope of the program and what it does. Distributing the application made it possible to run the application on different platforms and with different operating systems and to secure the compiled files, projects, and codes. So, this stage was the try- out stage for the COPTEFL system. It was passed over to the users to get feedback. Any bugs and glitches experienced during this stage were fixed.

## 2. Test administration:

This stage consists of two phases:

(a) **Try-out phase:** After the development of the software, the next step was to test it with users. 15 students who consented to participate were included in the trial. These were the students having their regular laboratory classes as part of their language program. In this phase, we gathered information on the usefulness of the test itself and for the improvement of the test and testing procedures.

(b) **Operational testing:** In this phase, the aim was to gather information on the usefulness of the test, but this time administering the test involved the goal to accomplish the specified use/purpose of the test. A total of 45 volunteer preparatory program students from various proficiency levels participated. One week before the administration, test-takers were divided into groups, each of which takes portions of the test at different times. One group was tested by the COPTEFL first, and then interviewed in the face-to-face speaking test, and a second group was provided the face-to-face speaking test first and then being tested by the COPTEFL. The test-takers who took the COPTEFL first were asked to take the face-to-face speaking test after 3 weeks and the test-takers who took the face-to-face speaking test first were asked to take the COPTEFL after 3 weeks. With a counterbalanced design like this, it was aimed to find out whether the test in one mode followed by the other could affect the score for the second mode. The following section described the step-by-step procedure of the program, the COPTEFL, and its administration:

**STEP 1:** The COPTEFL web page was loaded: [www.coptefl.com](http://www.coptefl.com)

STEP 2: Test-takers registered and logged in to the testing system.

STEP 3: Microphones were set up and tested.

STEP 4: Once the testing program started, the test-takers were presented with a short introductory page. In this introductory screen, the speaker welcomed the test-takers and introduced the test, providing information about its steps, format, function, procedure and length.

STEP 5: As soon as the test-takers listened to the instructions and clicked on the “next page” button, testing started.

STEP 6: Once the test-takers responded to all tasks, a final page was shown to express appreciation for taking the test. At this point, the program terminated and the test-takers exited the program.

After the administration of the tests, open-ended questions were given immediately in order to explore test-takers’ attitudes towards the test modes. After they completed writing, the researcher collected the answer sheets.

## **2.5. Data analysis**

### **2.5.1. Score comparability**

In order to evaluate the consistency between judges’ ratings, inter-rater reliabilities were calculated for each test delivery mode. To investigate the inter-rater reliability, a two-way random absolute agreement intra-class correlation coefficient (ICC) was performed. ICC assesses the consistency between judges’ ratings of a group of test-takers. Before proceeding to compare the magnitude of raw scores, the order effect on test scores was examined. To assess the effect of delivery mode and the mode-by-order interaction statistically, an Analysis of Covariance (ANCOVA) with within-subject effect was run on total scores.

For the investigation of the equivalence of test scores across modes, the magnitude of raw scores was compared by means of the two-way random absolute agreement intra-class correlation coefficient (ICC).

For better interpretations of the findings, mean score differences were also taken into account and therefore, mean scores by each test delivery mode were compared for total scores and task types. A paired samples t-test was run to compare the mean scores between the COPTEFL and the face-to-face speaking test.

### **2.5.2. Test-taker attitudes**

The answers of the test-takers were classified into themes each of which represented an idea related to their attitudes towards test conditions. After that, certain themes based on the ideas were coded and then placed into the categories each of which represented with. Finally, emerging themes were expressed in frequencies. In order to reduce research bias and establish the reliability of research findings, an expert from Anadolu University, who studies in the English Language Teaching Department as a research assistant and has experience in qualitative data analysis analyzed the data. To ensure the credibility of the findings, the consistency between the emerging findings from two researchers was investigated. Similarities and differences across the categories identified were sought out. After member checking sessions and rigorous discussions between the researchers, a final consensus on the categories was achieved. How well are the COPTEFL scores and the face-to-face speaking test scores compared in terms of (a) inter-rater reliability, (b) order-effect, and (c) test scores?



### 3. RESULT / FINDINGS

#### 3.1. Research Question 1

How well are the COPTEFL scores and the face-to-face speaking test scores compared in terms of (a) inter-rater reliability, (b) order-effect, and (c) test scores?

##### 3.1.1. Inter-rater reliability

ICC results for overall test scores and each task type test scores across delivery modes were provided in [Table 1](#) below.

**Table 1.** Intra-class correlation coefficient (ICC) results for inter-rater reliabilities across test delivery modes

Test score	<u>COPTEFL</u> ICC	<u>Face-to-Face</u> ICC
Overall score	.806*	.889*

\* significant at the .01 level

[Table 1](#) presented the results of the ICC on overall test scores for both delivery modes. As shown in the table, the ICC score for the face-to-face speaking test (ICC= .889) was slightly higher than the score in COPTEFL (ICC= .806). The average measure ICC for the COPTEFL was .806 with a 95% confidence interval from .64 to .89 ( $F(44,44)=5.075, p<.001$ ) and the average measure ICC for the face-to-face speaking test was .889 with a 95% confidence interval from .80 to .93 ( $F(44,44)=9.033, p<.001$ ). These ICC values between .75 and .90 indicated good reliability for both tests (Larsen-Hall, 2010). [Table 2](#) revealed the ICC results for each task type (see [Table 2](#)).

**Table 2.** Intra-class correlation coefficient (ICC) results of inter-rater reliabilities for each task type across test delivery modes.

Task type	<u>COPTEFL</u> ICC	<u>Face-to-face</u> ICC
Opinion	.686*	.796*
Comparison	.702*	.842*
Discussion	.772*	.890*

\* significant at the .01 level

For the COPTEFL, the findings showed that the average measure ICC for the opinion task was .688 with a 95% confidence interval from .42 to .82 ( $F(44,44)=3.151, p<.001$ ), for the comparison task, it was .702 with a 95% confidence interval from .45 to .83 ( $F(44,44)=3.317, p<.001$ ), and finally, for the discussion task, it was .772 with a 95% confidence interval from .58 to .87 ( $F(44,44)=4.320, p<.001$ ). As for the face-to-face speaking test, the findings showed that the average measure ICC for the opinion task was .796 with a 95% confidence interval from .63 to .88 ( $F(44,44)=4.943, p<.001$ ), for the comparison task, it was .842 with a 95% confidence interval from .71 to .91 ( $F(44,44)=6.260, p<.001$ ), and finally, for the discussion task, it was .890 with a 95% confidence interval from .79 to .94 ( $F(44,44)=8.913, p<.001$ ).

These results indicate a significant direct relationship between inter-rater reliability scores for each task type across test delivery modes that those who get higher scores from a task in the COPTEFL by the rater also get higher scores from the same task in the face-to-face speaking test by the inter-rater or vice versa.

For the COPTEFL, the findings revealed that the ICC values for opinion and comparison tasks were between .50 and .75, meaning that the level of reliability was moderate. For the discussion

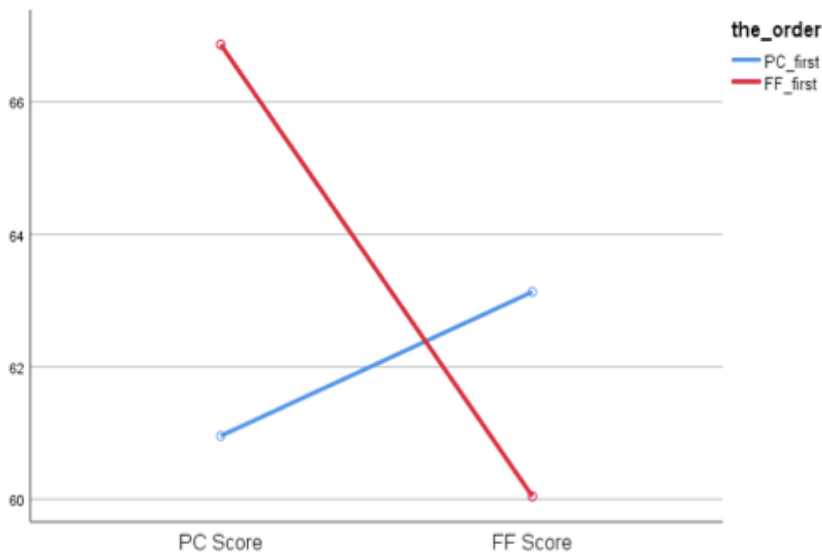
task, it was .77, which indicated good reliability. As for the face-to-face speaking test, each ICC value was between .75 and .90, revealing that the level of reliability was good.

### 3.1.2. Order-effect

Prior to comparing raw scores awarded to the two delivery modes, an Analysis of Covariance (ANCOVA) was computed to assess whether the existence of order has an effect on speaking test scores. The findings revealed that there is a significant interaction between test order and test scores ( $F(1,43)=6.89, p=.012$ ).

As Figure 12 shows, the groups which took the COPTEFL first did better on the face-to-face speaking test ( $M=63.13, SD=11.83$ ) than on the COPTEFL ( $M=60.96, SD=11.82$ ); and the groups which took the face-to-face speaking test first did better on the COPTEFL ( $M=66.86, SD=11.77$ ) than on the face-to-face speaking test ( $M=60.05, SD=11.39$ ) independent of their level.

**Figure 12.** The interaction between group and test mode.



The results suggest that both groups did better in their second test than in their first test no matter which type of test they took first, which shows that there was a practice effect in general.

### 3.1.3. Comparability of raw scores

The analyses in this part focused on the differences in test scores between the COPTEFL and the face-to-face speaking test. In order to determine differences, ICC and paired samples t-tests were computed across delivery modes.

#### a. The relationship between scores by test delivery modes:

A two-way random absolute agreement ICC was conducted in order to find out how well the COPTEFL scores and the face-to-face speaking test scores were correlated. The analyses were run for total scores and task type scores across test delivery modes (see Table 3).

A significant moderate degree of ICC was found between the COPTEFL total scores and face-to-face speaking test total scores. The average measure ICC was .632 with a 95% confidence interval from .33 to .79 ( $F(44,44)=2.735, p<.001$ ). This result indicates a direct relationship between the scores across test delivery modes that those who get higher scores from COPTEFL also get higher scores from the face-to-face speaking test, or vice versa.

**Table 3.** ICC results for total scores and task type scores across test delivery modes.

	Total score	COPTEFL		
		Opinion task	Comparison task	Discussion task
<u>Face-to-face</u>				
Total score	.632*			
Opinion task	-	.382		
Comparison task	-	-	.576**	
Discussion task	-	-	-	.704*

\* significant at the .01 level

\*\* significant at the .05 level

As for the scores from each task type, a low degree of ICC was found between the COPTEFL opinion task scores and the face-to-face speaking test opinion task scores. The average measure ICC was .382 with a 95% confidence interval from -.09 to .65 ( $F(44,44)=1.648, p>.05$ ). For the comparison task scores, there was a significant moderate degree of absolute agreement ICC between the COPTEFL and the face-to-face speaking test. The average measure ICC was .576 with a 95% confidence interval from .22 to .76 ( $F(44,44)=2.347, p<.05$ ), which indicates that those who get higher scores from the comparison task in the COPTEFL also get higher scores from the comparison task in the face-to-face speaking test, or vice versa. Finally, as for the discussion task scores, the results revealed a significant moderate degree of ICC between the COPTEFL and the face-to-face speaking test. The average measure ICC was .704 with a 95% confidence interval from .45 to .83 ( $F(44,44)=3.333, p<.001$ ), which shows that those who get higher scores from the discussion task in the COPTEFL also get higher scores from the discussion task in the face-to-face speaking test, or vice versa.

b. Comparing the mean scores by test delivery modes:

For better interpretations of the findings, mean score differences were also taken into account and therefore, mean scores by each test delivery mode were compared for total scores and task types. A paired samples t-test was run to compare the mean scores between the COPTEFL and the face-to-face speaking test (see Table 4).

**Table 4.** Paired samples t-test results for each task type across test delivery modes.

	COPTEFL		Face-to-face		df	t	p
	Mean	SD	Mean	SD			
Total score	63.84	12.04	61.62	11.59	44	1.219	.229
Opinion task	65.36	13.57	61.07	12.28	44	1.808	.077
Comparison task	63.80	12.66	62.27	12.68	44	0.743	.462
Discussion task	62.82	13.85	62.33	12.57	44	0.258	.798

As Table 4 presented, the findings showed that there is not a statistically significant difference between the COPTEFL and the face-to-face speaking total test scores ( $t(44)= 1.219; p>.05$ ). When the mean scores of each task type were investigated, the findings showed that there is not a statistically significant difference between the COPTEFL scores and the face-to-face speaking test scores for task types, which are the opinion task ( $t(44)= 1.808; p>.05$ ); the comparison task ( $t(44)= 0.743; p>.05$ ), and the discussion task ( $t(44)= 0.258; p>.05$ ). Therefore, it can be concluded that test delivery mode was found not to have a significant effect on test-takers' speaking test scores.

### 3.2. Research Question 2

What are the attitudes of test-takers in relation to the test delivery modes?

### 3.2.1. General attitudes towards the COPTEFL

To explore the face validity of the COPTEFL from the test-takers' perspective, an open-ended question assessing participants' attitudes towards the COPTEFL was posed. The findings are presented in order of frequency below (see Table 5).

**Table 5.** Test-takers' general attitudes towards the COPTEFL.

Positive comments on COPTEFL		Negative comments on COPTEFL	
Categories	Num.**	Categories	Num.**
1. Test system	42	1. Test system	11
User friendly		Weak microphones	
Well designed		Abrupt transition between	
Easy to start and follow		instructions and tasks	
Easier to understand the pronunciation		The effect of the countdown	
Practical		timer on the screen	
Good sound quality			
2. Tasks	10	2. Tasks	2
At the right level of difficulty		Tough questions	
3. Time limit	11	3. Time limit	13
Enough time to give answers		Little answering time	
Enough time to think about answers		Little thinking time	

\*\*The number of the comments by test-takers

Nearly all of the participants stated that the test system was well designed, quite easy to operate, and works well:

*"The COPTEFL system was quite easy to access and operate."* P3

*"I think the COPTEFL is much more practical than the face-to-face speaking test. It does not require teachers to interview and this saves time for them."* P12

According to them, the system was user friendly and provided practical experience for test-takers and teachers. Some of them reported that the sound quality was satisfactory and they had no difficulty in hearing or understanding the instructions or the questions:

*"In face-to-face speaking tests, it is sometimes difficult to understand the interviewer's pronunciation. In the computerized test, on the other hand, the correctness of pronunciation was controlled beforehand, and the questions were shown on the screen. This made the tasks clear to understand."* P8

Although most of the comments were positive in relation to the test system, there were a few constructive comments for improvement of the system or the tasks:

*"It would be better if the time limit for speaking was longer."* P16

*"Tasks were tough, so the number of tasks could be reduced or the time limit could be multiplied."* P24

In some of the comments, test-takers stated that microphones could be better in order to achieve the best possible results for sound recording. Also, some of them referred to the bad effects of seeing countdown timer on the screen:

*"Countdown-timer made me feel nervous."* P7

Apart from those, one of the participants also made a comment on the transition from instructions to tasks. According to her, that transition was abrupt and due to this, she got anxious.

### 3.2.2. The direct comparison of two modes

To better understand test-takers' test method preferences, the responses to the second open-ended question were analyzed. Of those analyses, three categories were developed based on the comments that favored the face-to-face mode. These were presented in order of frequency below (Table 6).

**Table 6.** The direct comparison of two modes.

Attitudes towards the face-to-face speaking test		Attitudes towards the COPTEFL	
Categories	Num.**	Categories	Num.**
1. Interaction	15	1. Less anxiety	29
2. Naturalness	4	2. Better control	4
3. More time	1	3. Test fairness	2

\*\* The number of the comments by test-takers

The main reason test-takers had more favorable attitudes to the face-to-face speaking test was the interaction with the interviewer. The participants remarked that they performed better on this test since the reactions from the interviewer such as smiling and nodding helped them feel comfortable and relaxed. Although the interviewers did not assist, test-takers were still trying to figure out whether or not they were being understood thanks to the facial expressions of the interviewers. According to them, non-verbal communication should be involved in an examination atmosphere, because no reactions could make them unable to gauge how far they came to the correct answer:

*“During the exam, I would prefer to have feedback from the teacher to get certain about the correctness of my responses. So, I would prefer face to face speaking exams.” P17*

*“Interaction with the teacher helped me speak more.” P3*

Some of the participants stated that it felt more natural to talk in the presence of the interviewer since it was similar to a real-life conversation where the communication is between two or more people. Two of them perceived the face-to-face speaking test as a better measure of their spoken English because of this sense of naturalness. Although some of the test-takers preferred having a conversation with an interviewer who could accommodate their responses and the use of time, the opposite was also true for some others who preferred the COPTEFL due to lack of influence of the interviewer and the use of time:

*“I got nervous in face-to-face exams. But, the COPTEFL made me feel relaxed since I was testing myself alone.” P3*

*“There would be no influence of the interviewer who was faced with a problem just before the exam and reflected it on us.” P8*

*“Sometimes the way interviewers behave in the test makes me nervous. But, in the COPTEFL, there is not such a problem. P15*

In conclusion, the quantitative data showed that the test-takers performed better on the COPTEFL compared to the face-to-face speaking test ( $M= 63.84$ ,  $M= 61.62$ , respectively). The qualitative data provided insights into the attitudes of test-takers in relation to both test modes and revealed that if given choice, many of them preferred the face-to-face speaking test due to the opportunity of interaction with the interviewer while some of them have a strong preference for the COPTEFL due to its provoking less anxiety. These results showed that different types of learners have different testing experiences and thus preferred either the COPTEFL or the face-to-face speaking test.



#### 4. DISCUSSION and CONCLUSION

The subjectivity in the rater judgments is one of the major sources of measurement error and a threat to the reliability and validity of test scores (Bachman, Lynch & Mason, 1995). Inter-rater reliability estimates in the present study were .80 (ICC) for the computer mode and .88 (ICC) for the face-to-face mode, indicating that the level of reliability for each mode was good. One possible interpretation of this result is that a potential problem of inconsistency in different raters' scores was effectively controlled. The issue of experience at this point is considered as "the most important reason for rating scales appearing to be meaningful and providing reliable results" (p.97). In the present study, experts in testing speaking rated to an existing scale developed by a formal testing body in AUSFL. In order to achieve a common standard –which no one would wish to disagree with, rater training and socialization into the use of the scale were valued in the study. Such training was perceived as the way to ensure greater reliability and validity of scores produced in language performance tests (Fulcher, 2014). With rater training sessions, it was intended to "socialize raters into a common understanding of the scale descriptors, and train them to apply these consistently in operational speaking tests" (Fulcher, 2014, p.145). These efforts could possibly lead to the achievement of good inter-rater reliability scores for both tests. Having two raters instead of one might also be one of the reasons for achieving good reliability. As argued in Fulcher (2014), the use of a double rating can avoid the potential effect that an individual rater may have on the test score. That is, multiple ratings of each performance help minimize the subjectivity of ratings and therefore, improve reliability (Carr, 2011).

Both groups did better on their second test than on their first test no matter which test mode they took first. This finding revealed that there was a practice effect in general. The question then arises as to why there was a practice effect. One possible explanation lies in that the group who took the COPTEFL first performed poorly in the COPTEFL mode because they had never taken a speaking test in a CBT mode and therefore, might not achieve the best performance due to their unfamiliarity with the test format. When provided the opportunity to take a second test, they might demonstrate better performance. Similar to those who took the COPTEFL first, test-takers who took the face-to-face speaking test first might achieve a higher score on the second test since they became familiar with the test content. In line with this finding, Öztekin (2011) also reported a test order effect in her study results in which both groups did better on their second test than on their first test no matter which type of test they took first. Similarly, Zhou (2009) revealed that the test order effect was present in the findings of the study. But, this time, the group who took the computer-based speaking test first performed better on the later face-to-face speaking test, whereas the group who took the face-to-face speaking test first did not perform better on the computer-based test. In relation to this finding, Zhou (2009) states that the reason behind this finding may lie perhaps in the reactions from the interviewer. Accordingly, during the face-to-face speaking test, the reactions from the interviewer might have motivated the test-takers to give better verbal responses to the tasks and therefore they may have felt encouraged to do their best by the presence of the interviewer.

The lack of statistically significant differences between the mean scores indicated that test-takers who did well on the COPTEFL did almost equally well on the face-to-face speaking test and there was no major change in test-takers' performance on monologic speaking tasks when the response was elicited through non-human elicitation techniques. This finding suggested that test delivery mode did not account for the variance in test scores in the present study. With regard to the comparisons between the computer mode and the face-to-face mode, some studies have also shown a considerable overlap between delivery modes for speaking tests, at least in the correlational coefficient sense that test-takers who score high in one mode also score high in the other or vice versa (Mousavi, 2007; Thompson, Cox & Knapp, 2016; Zhou, 2015). This

research was correlational in nature and the correlation coefficient between the test modes was found to be .63, which is considered a moderate index of reliability. In conformity with the traditional requirements for concurrent validation (Alderson, Clapham & Wall, 1995), a correlation coefficient of .9 or higher is indicated to be the appropriate level of standards at which test users could consider “the semi-direct testing results closely indicative of probable examinee performance on the more direct measures” (Clark, 1979, p.40). Even though the correlation coefficient score in the present study was lower than the figure of .9, as Mousavi (2007) put forward for a figure of .63, the finding highlighted the usability of the newly developed prototype test of oral proficiency as a reasonable alternative mode of test delivery.

One possible explanation of the findings in the study might be that test-takers performed similarly in both test delivery modes and small differences between individual scores that are statistically non-significant might not be detected by using the paired-sample t-test. As Kiddle and Kormos (2011) report, the correlational analysis measures the strength of the relationship between the two delivery modes and high correlations might be accomplished even if the test-takers score differently in the two modes. If, for example, test-takers were consistently awarded higher scores in the face-to-face mode than in the computer mode, correlations can still remain high. At this point, Kiddle and Kormos (2011) suggest further empirical analysis such as Rasch analysis that

“Unlike analyses such as t-tests and correlations, the Rasch analysis does not rely on raw scores but uses logit scores instead, and consequently can yield reliable information on whether the fact that the test was administered under different conditions has an effect on test performance” (p.353).

As for the interpretation of the lack of significant differences in the mean scores across modes, one of the reasons might be that test-takers performed differently between two modes, but raters tended to award similar scores to the test-takers across tasks based on their overall impression about them on a particular task or the overall test. Gülle (2015), for example, pointed out that raters might show a tendency to assign similar scores to the test-takers across tasks due to their holistic judgments. In the current study, in order to minimize possible halo effects, the raters were assigned the scoring criteria for each task separately. Instead of rating one test-taker on all three tasks, they were asked to award scores for the test-takers’ performances on the first task and then continue with the second task and the third task. However, as Gülle (2015) states, it is still possible for the raters to assign similar scores across different tasks based on their overall judgments of the test-taker performances. The present results may also be attributed to the possibility that test-takers performed differently between two tests, but not to extent that the raters were able to discern due to the little difference between bands on a given subscale.

Qualitative analysis of the data revealed that test-takers had favorable attitudes towards the COPTFL in many aspects and the majority of them did not show a particular preference in terms of the testing modes. But, it appears from the responses, if given the choice, most of the test-takers were found to prefer the face-to-face speaking test. However, this finding did not necessarily imply that their reactions to the COPTFL were negative. This finding corroborated with the finding of Qian (2009) who also found that participants did not have a particular preference with respect to the testing modes, and only partially corroborated with the findings of most researchers including McNamara (1987); Shohamy, Donitsa-Schmidt, and Waizer (1993) and Joo (2008), who all found that an overwhelming of participants showed a particular preference to the direct testing mode. The finding of the current study was at odds with Brown’s (1993) study that test-takers preferred semi-direct testing mode to the direct testing one. At this point, as Qian (2009) suggested that we should

“be cautious about drawing a conclusion as to which testing mode is more amenable to test-takers as their preferences might be test and context dependent: Test-takers’ attitudes may be

influenced by various factors, such as test quality, the stakes of the test to the test-taker, test-takers' cultural traditions and personalities, and so forth" (p.123).

#### 4.1. Limitations of the study

The findings of the study must be considered in the context of several potential limitations and therefore, some caution is warranted when interpreting and generalizing the study results. It must be noted that the COPTEFL, as a newly developed testing format, was conducted on a small scale with a relatively small number of test-takers. The sample size which was drawn only from AUSFL might not have been representative or provided sufficient data for the study. Also, it should be stated that the sample only included AUSFL students who were mostly pre-intermediate or intermediate level students having a similar educational background and high computer familiarity due to their classes in the laboratories as a part of their regular language program. Thus, in more diverse and larger sample size, a more convincing and distinct or even different set of results might have been arrived at. The conclusions drawn from the analyses were, therefore, tentative.

#### 4.2. Implications of the study

The findings of the study suggested that before serving the COPTEFL as a substitute for the face-to-face speaking tests, making students familiar with it through several practices and therefore, helping them get used to is important. So, during these practice sessions, test administrators should explain test-takers the differences in testing formats and take their preferences into consideration, and thus support them when selecting the testing format that best meets their needs and interests. When test-takers get used to the COPTEFL, they can benefit from it as a new learning experience. On successful administration of it, the COPTEFL can be administered in language laboratory classes where students practice their English. This can help students assess their language on their own and see the progress in their level of English in time since the sounds are recorded. In time, practicing with the COPTEFL may help students to reduce the levels of nervousness mostly associated with face-to-face speaking tests. By giving award scores to their students constantly, the teachers can be informed about the profile of their students during the education period. The results of the experiment also showed that the use of the COPTEFL as a testing format helped reduce the amount of time, human resources, number of proctors, space and hard copy material required for a face-to-face speaking test. In addition to its advantages to test administrators and language testers, raters can also benefit from the convenience and user-friendliness of the rating platform in the COPTEFL system, which is attractive for its availability at any time and any place, low cost (i.e. no need for the use of cameras and CDs and a technical team to set the cameras) and potentially increased effectiveness compared with traditional face-to-face delivery. In foreign language programs at universities where there are a large number of incoming international students or exchange students (i.e. Erasmus students), it is generally a problem to group newcomers into appropriate instructional levels due to restricted time for organizing and delivering an entry or a placement test. For such cases, the COPTEFL can serve as a standardized oral proficiency test.

The issues in the development of the COPTEFL have also important implications for future computer-based speaking test developers. The step-by-step processes in the test design, construction and administration can be used as a roadmap for the test developers. This study can only be considered a first approach for a computer-based speaking test. Future researchers can improve on this study by changing the nature of task types, the number of tasks or the scoring system. Finally, the limited number of research of this type in Turkey, also, may be a reason to further study in this field.

## Acknowledgements

This study is a part of the Doctoral Thesis of Cemre İşler in the Program of English Language Teaching in Anadolu University.

## Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

## Authorship contribution statement

**Cemre Isler:** Investigation, Resources, Software Development, Data Collection, Data Analysis, and Writing the original draft. **Belgin Aydın:** Methodology, Supervision and Validation.

## ORCID

Cemre İşler  <https://orcid.org/0000-0002-3622-0756>

Belgin Aydın  <https://orcid.org/0000-0002-4719-7440>

## 5. REFERENCES

- Alderson, J.C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.
- Aydın, B., Akay, E., Polat, M., & Geridönmez, S. (2016). Türkiye’deki hazırlık okullarının yeterlik sınavı uygulamaları ve bilgisayarlı dil ölçme fikrine yaklaşımları. *Anadolu Üniversitesi Sosyal Bilimler Dergisi*, 16(2), 1-20.
- Aydın, B., Geridönmez, S., Polat, M. & Akay, (2017). Feasibility of computer assisted English proficiency tests in Turkey: A field study. *Anadolu Üniversitesi Eğitim Bilimleri Enstitüsü Dergisi*, 7(1), 107-122.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.
- Bachman, L. F., Lynch, B., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, 12(2), 238-57.
- Carr, N. T. (2011). *Designing and analyzing language tests: Oxford handbooks for language teachers*. Oxford University Press.
- Chapelle, C. A. (2013). Conceptions of validity. In *The Routledge handbook of language testing* (pp. 35-47). Routledge.
- Clark, J. L. D. (1979). Direct vs. semi-direct tests of speaking ability. In E. J. Briere & F. B. Hinofotis (Eds.), *Concepts in language testing: some recent studies*, 35-49. TESOL.
- Council of Europe, (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.
- Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative*. Pearson Education.
- East, M. (2016). Mediating Assessment Innovation: Why Stakeholder Perspectives Matter. In *Assessing Foreign Language Students’ Spoken Proficiency* (pp. 1-24). Springer.
- Field, J. (2013). Cognitive Validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening*, *Studies in language testing*, 35 (pp.77-151). Cambridge University Press.
- Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing*, 20(4), 384-408.
- Fulcher, G. (2014). *Testing second language speaking*. Routledge.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. Routledge.



- Galaczi, E. D., & French, A. (2011). Context validity of Cambridge ESOL speaking tests. In L. Taylor (Ed.), *Examining speaking*, 30. Cambridge University Press.
- GSE. (2015). *New Global Scale of English Learning Objectives*. Pearson English. Retrieved May 18, 2017, from <http://www.english.com/gse#>
- Gülle, T. (2015). *Development of a speaking test for second language learners of Turkish* [Unpublished master's thesis]. Boğaziçi University.
- Jeong, T. (2003). *Assessing and interpreting students' English oral proficiency using d-VOCI in an EFL context* [Unpublished doctoral dissertation]. Ohio State University, Columbus.
- Joo, M. (2008). *Korean university students' attitudes to and performance on a Face-To-Face Interview (FTFI) and a Computer Administered Oral Test (CAOT)* [Doctoral dissertation]. University of London.
- Kenyon, D. M., & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning and Technology*, 5(2), 60-83.
- Kiddle, T., & Kormos, J. (2011). The effect of mode of response on a semidirect test of oral proficiency. *Language Assessment Quarterly*, 8(4), 342-360.
- Khalifa, H., & Weir, C. J. (2009). *Examining reading* (Vol. 29). Ernst Klett Sprachen
- Larsen-Hall, J. (2010). *A guide to doing statistics in second language research*. Routledge.
- Luther, A. C. (1992). *Designing interactive multimedia*. Multi-science Press Inc.
- Malabonga, V., Kenyon, D. M. & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing*, 22(1), 59-92.
- Messick, S. (1989) Validity. In R. L., Linn (Eds.), *Educational measurement* (pp. 13-103). Macmillan/American Council on Education.
- McNamara, T. F. (1987). *Assessing the language proficiency of health professionals. Recommendations for the reform of the Occupational English Test (Report submitted to the Council of Overseas Professional Qualifications)*. Department of Russian and language Studies, University of Melbourne, Melbourne, Australia.
- Mousavi, S. A. (2007). *Development and validation of a multimedia computer package for the assessment of oral proficiency of adult ESL learners: implications for score comparability* [Doctoral dissertation]. Griffith University.
- Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal*, 93(1), 836-847.
- O'Loughlin, K. (1997). *The comparability of direct and semi-direct speaking tests: A case study* [Unpublished doctoral dissertation]. University of Melbourne.
- O'Sullivan, D. B. (Ed.). (2011a). *Language testing: Theories and practices*. Palgrave Macmillan.
- O'Sullivan, B. (2011b). Language testing. In *The Routledge handbook of applied linguistics* (pp. 279-293). Routledge.
- Öztekin, E. (2011). *A comparison of computer assisted and face-to-face speaking assessment: Performance, perceptions, anxiety, and computer attitudes* [Master's thesis]. Bilkent University.
- Qian, D. (2009). Comparing direct and semi-direct modes for speaking assessment: affective effects on test takers. *Language Assessment Quarterly*, 6(2), 113–125.
- Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing* (Vol. 26). Cambridge University Press.
- Shneiderman, B. (2004). *Designing the user interface: Strategies for effective human-computer interaction* (4th edition). Addison-Wesley.
- Shohamy, E., Donitsa-Schmidt, S., & Waizer, R. (1993). *The effect of the elicitation mode on the language samples obtained in oral tests* [Paper presentation]. 15th Language Testing Research Colloquium, Cambridge, UK.

- Thompson, G. L., Cox, T. L., & Knapp, N. (2016). Comparing the OPI and the OPIc: The effect of test method on oral proficiency scores and student preference. *Foreign Language Annals*, 49(1), 75-92.
- Weir, C. J. (2005). *Language testing and validation*. Palgrave MacMillan.
- Wu, W. M., & Stansfield, C. W. (2001). Towards authenticity of task in test development. *Language Testing*, 18(2), 187-206.
- Zainal Abidin, S. A. (2006). *Testing spoken language using computer technology: A comparative validation study of 'Live' and computer delivered test versions using Weir's framework* [Doctoral dissertation]. Universiti Teknologi Mara.
- Zak, D. (2001). *Programming with Microsoft Visual Basic 6.0: Enhanced edition*. Course Technology Thomson Learning.
- Zhou, Y. (2008). A comparison of speech samples of monologic tasks in speaking tests between computer-delivered and face-to-face modes. *JLTA Journal*, 11, 189-208.
- Zhou, Y. (2009). *Effects of computer delivery mode on testing second language speaking: The case of monologic tasks* [Doctoral dissertation]. Tokyo University of Foreign Studies.
- Zhou, Y. (2015). Computer-delivered or face-to-face: Effects of delivery mode on the testing of second language speaking. *Language Testing in Asia*, 5(1), 2.