# DEVELOPING AND VALIDATING A COMPUTERIZED ORAL PROFICIENCY TEST OF ENGLISH AS A FOREIGN LANGUAGE (COPTEFL)

**Doktora Tezi**

**Cemre İŞLER**

**Eskişehir 2019**

# DEVELOPING AND VALIDATING A COMPUTERIZED ORAL PROFICIENCY TEST OF ENGLISH AS A FOREIGN LANGUAGE (COPTEFL)

**Cemre İŞLER**

**PhD DISSERTATION**

**Department of Foreign Language Education**

**Supervisor: Prof. Dr. Belgin AYDIN**

**Eskişehir**

**Anadolu University**

**Graduate School of Educational Sciences**

**July 2019**

JÜRİ VE ENSTİTÜ ONAYI

Cemre İŞLER'in "Developing and Validating A Computerized Oral Proficiency Test of English As A Foreign Language (COPTEFL)" başlıklı tezi 23.05.2019 tarihinde aşağıdaki jüri tarafından değerlendirilerek "Anadolu Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği"nin ilgili maddeleri uyarınca Yabancı Diller Eğitimi Anabilim Dalı İngilizce Öğretmenliği Programında, Doktora tezi olarak kabul edilmiştir.

| | Unvanı-Adı Soyadı | İmza |
|---|---|---|
| Üye (Tez Danışmanı) | : Prof.Dr. Belgin AYDIN | |
| Üye | : Prof.Dr. Paşa Tevfik CEPHE | |
| Üye | : Doç.Dr. Özgür YILDIRIM | |
| Üye | : Doç.Dr. Murat AKYILDIZ | |
| Üye | : Dr. Öğr. Üyesi Aylin ÜNALDI | |

Doç.Dr. Yasemin ERGENEKON
Anadolu Üniversitesi
Eğitim Bilimleri Enstitüsü
Müdür Vekili

**ABSTRACT**

DEVELOPING AND VALIDATING A COMPUTERIZED ORAL PROFICIENCY TEST OF ENGLISH AS A FOREIGN LANGUAGE (COPTEFL)

Cemre İŞLER

Department of Foreign Language Education, Programme in English Language Teaching

Anadolu University, Graduate School of Educational Sciences, July 2019

Supervisor: Prof. Dr. Belgin AYDIN

This study is about the development and validation process of the Computerized Oral Proficiency Test of English as a Foreign Language (COPTEFL). 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. Following Weir's (2005) socio-cognitive framework, the study provided evidence for cognitive validity along with context and scoring validity aspects of the tests. After the development of the tasks, it was aimed to develop the COPTEFL system and then compare the test scores on monologic tasks between the COPTEFL and the face-to-face speaking test. The study also explored test comparability from the test takers' perspectives and compared the test taker attitudinal reactions to taking different formats. The study was employed at AUSFL, Turkey. 45 students participated in the study. To investigate the test score comparability, intra-class correlations coefficients (ICC) and paired-samples *t*-tests were run. For test taker attitudes, qualitative content analysis scheme of Creswell (2012) was employed. The findings from these quantitative and qualitative analyses provided substantial support for the validity and reliability of the COPTEFL and inform the further refinement of the test tasks. In the light of discussions based on the results of the study, suggestions for further research were also mentioned.

**Keywords:** Language testing, Oral proficiency testing, Computerized oral proficiency testing, EFL context.

# ÖZET

BİLGİSAYAR TABANLI İNGİLİZCE KONUŞMA YETERLİK TESTİ GELİŞTİRİLMESİ VE GEÇERLİK ÇALIŞMASI (COPTEFL)

Cemre İŞLER

Yabancı Diller Eğitimi Anabilim Dalı, İngilizce Öğretmenliği Programı

Anadolu Üniversitesi, Eğitim Bilimleri Enstitüsü, Temmuz 2019

Danışman: Prof. Dr. Belgin AYDIN

Bu araştırma, Bilgisayar Tabanlı İngilizce Konuşma Yeterlik Testi'nin (COPTEFL) geliştirilmesi ve geçerlilik çalışması sürecine yönelik bir çalışmadır. Testin amacı, Anadolu Üniversitesi Yabancı Diller Yüksekokulu (AUSFL) öğrencilerinin konuşma yeterlilik seviyelerini ölçmektir. Bu amaçla, Evrensel Dil Ölçeği (GSE, 2015) seviye tanımlayıcıları temel alınarak üç farklı konuşma testi sorusu hazırlanmıştır. Çalışma, Weir'in (2005) sosyo-bilişsel çerçevesini esas alarak bilişsel, bağlamsal ve puanlama geçerliliği yönlerinden testin geçerliliğine kanıt sağlamıştır. Test sorularının geliştirilmesinden sonra, COPTEFL sisteminin geliştirilmesi ve ardından COPTEFL ile yüz yüze konuşma testi arasındaki test puanlarının karşılaştırılması hedeflenmiştir. Çalışmada ayrıca, katılımcıların test ortamlarına yönelik tutumları da incelenmiştir. Araştırma, AUSFL, Türkiye'de uygulanmış ve çalışmaya, 45 öğrenci katılmıştır. Test puanlarının karşılaştırılabilirliğini araştırmak için, Sınıf İçi Korelasyon katsayıları (ICC) ve Eşleştirilmiş Örneklem t- Testleri yapılmış, teste girenlerin tutumlarını araştırmak içinse Creswell'in (2012) nitel içerik analizi yöntemi kullanılmıştır. Bu nitel ve nicel analizlerden elde edilen bulgular, COPTEFL'ın geçerliliği ve güvenirliği konusunda önemli kanıtlar sunmuştur. Çalışmanın sonucu baz alınarak yapılan tartışmaların ışığında, gelecek araştırmalar için önerilerden de bahsedilmiştir.



**Anahtar Sözcükler:** Yabancı dilde ölçme, Konuşma yeterlik testi, Bilgisayar tabanlı konuşma yeterlik testi, İngilizce'nin yabancı dil olarak öğretildiği ortam.

## ACKNOWLEDGEMENT

This study could not have been completed without the cooperation of a large number of people. I am deeply indebted to my supervisor, Prof. Dr. Belgin AYDIN whose feedback, stimulating suggestions and encouragement helped me in all the time of research for and writing of this thesis. Many special thanks to Assoc. Prof. Dr. Özgür YILDIRIM for his continuous support and Assoc. Prof. Dr. Murat AKYILDIZ for his meticulous feedback. I am also grateful to the other committee members, Prof. Dr. Paşa Tevfik CEPHE and Asst. Prof. Dr. Aylin ÜNALDI for their acceptance to be in my jury and their constructive feedback. My heartfelt thanks go to my brother Yücel TACAL who supported me during the software development process. I am also grateful to all Anadolu University School of Foreign Languages (AUSFL) testing unit and to the students who volunteered to undergo up to several times language testing. Besides, I would like to extend my special thanks to my colleague, Tuncay KARALIK, who accompanied me analyzing the qualitative data.

Additionally, I would like to thank TÜBİTAK for the scholarship they provided me from the beginning of my PhD.

Above all these, I wish to express my sincerest gratitude to my family for their endless patience, care and love.

Cemre İŞLER

ESKİŞEHİR 2019

**04.07.2019**

## STATEMENT OF COMPLIANCE WITH ETHICAL PRINCIPLES AND RULES

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

**Cemre İŞLER**

## TABLE OF CONTENTS

Page

Page

## LIST OF TABLES

## LIST OF FIGURES

**Page**

## LIST OF ABBREVIATIONS

| | |
|---|---|
| AUSFL | : Anadolu Unıversity School of Foreign Languages |
| CBT | : Computer-based Testing |
| CEFR | : Common European Framework of Reference |
| CLA | : Communicative Language Ability |
| COPTEFL | : Computerized Oral Proficiency Test of English |
| EFL | : English as a Foreign Language |
| GSE | : Global Scale of English |
| OPI | : Oral Proficiency Interview |
| SOPI | : Simulated Oral Proficiency Interview |

# 1. INTRODUCTION

## 1.1. Background to the Study

Language teachers devote considerable time and energy to their learners in classrooms. During teaching processes, they are continually evaluating their learners' performance in the second language, noting whether or not their communication is interfered by defective grammar or pronunciation (Madsen, 1983; Douglas, 2014). Almost unaware, they are administering little oral assessment as they offer feedback informally to learners about their language ability. As Madsen (1983) states such evaluation happens very naturally in the classrooms even during the early weeks of the class. If teachers can assess their learners' language ability like this – and most of the teachers are called to do so by the early months of instruction - why do they need language tests? Probably, the most important reason is fairness (Douglas, 2014). Tests make it possible for teachers to ensure that they treat all their students the same by providing each of them an equal opportunity to show what they can do with the language (Douglas, 2014). No student can be defined as being proficient in a language just because the teacher's presumed and sometimes real biases are an issue for what students have learned and what they can do with the language they have learned. Tests also make it possible for teachers to make decisions about students' needs with more confidence. They offer some standardization by which teachers can monitor performance and progress, allowing them to compare students with each other or against some external performance criteria. It can therefore be revealed that language ability –a person's ability to express facts, ideas, feelings and attitudes clearly and with ease, in speech or in writing, and the ability to understand what he or she hears and reads (Heaton, 1990) –can best be measured by tests which evaluate performance in the language skills. Listening and reading comprehension tests, oral interviews, and letter writing are used to test ability in those language skills performed in real life. Ways of measuring performance in the four skills may take the form of tests of (Heaton, 1990, p.8):

i. Listening (auditory) comprehension, in which short utterances, dialogues, talks and lectures are given to the test takers;

ii. Reading comprehension, in which questions are designed to test the learners' ability to get the gist of a text and to extract central information on specific points in the text;
iii. Writing ability, generally in the form of letters, reports, memos, messages, instructions, and accounts of past events;
iv. Speaking ability, generally in the form of an interview, a picture description, role-play, and a problem-solving task involving pair work or group work, etc.

Among the four skills, the testing of speaking is widely considered as the most challenging of all language proficiency exams to prepare, administer, and score (Madsen, 1983). What are some of the reasons for speaking test's being so challenging? One reason is that the complex nature of speaking makes it difficult to measure accurately (Madsen, 1983; Fulcher, 2003). There is a clear lack of understanding of what constitutes speaking as construct. Grammar, vocabulary, and pronunciation are usually called as ingredients. But areas such as fluency and appropriateness of expression are often considered as equally important. In addition, there is also not very wide agreement on how to weight each factor (such as fluency or grammar) (Madsen, 1983). Apart from all of these factors, there is also the practical problem of having to test each student's oral proficiency individually (Malabonga, Kenyon & Carpenter, 2005). That is, the administration of oral proficiency testing is a time consuming and labor-intensive process. 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 the speaking skill.

There are a number of approaches to oral proficiency testing, including direct, and indirect or namely semi-direct speaking tests. Direct approaches involve assessments carried out by a live interviewer who elicits language from the test taker via a face-to-face or telephonic oral proficiency interview. The American Council on the Teaching of Foreign Languages (ACTFL) has conducted thousands of the Oral Proficiency Interviews (OPIs) to test the oral proficiency of students and professionals in the United

States over the past 25 years. In an ACTFL OPI, a well-trained and certified interviewer elicits a speech sample through a structured yet flexible interview constructed to measure the test taker's proficiency level according to the ACTFL Speaking Proficiency Guidelines (Malabonga, Kenyon & Carpenter, 2005). But face-to-face testing of oral proficiency can be expensive and time consuming, which usually makes indirect or semi-direct testing more attractive to test users (Malone, 2007).

Indirect speaking test, on the other hand, has no face-to-face interaction with an interlocutor, and is equivalent to the use of 'semi-direct' in some of the literature (Fulcher, 2003). Semi-direct approaches refer to testing methods that rely on something other than a live interviewer—such as a tape recorder to elicit language from the test taker. The Simulated Oral Proficiency Interview (SOPI) is a tape-mediated speaking test. It follows the general structure of the OPI used by government agencies and the ACTFL to test speaking proficiency. While the OPI is a face-to-face interview, the SOPI relies on audiotaped instructions and a test booklet to elicit language from the test taker.

In the OPI, test takers have a certain amount of input into the procedure. The interviewer adapts the level of difficulty of the questions to the proficiency level displayed by the test taker. Test takers control the length of their responses to the interviewer's questions. They have some control over the content of the interview in that the interviewer is trained (particularly at lower levels) to follow up on topics nominated by the test taker. In tape-mediated tests, much of this control is lost. In general, timed pauses prescribe for the test taker how much time he or she has to think about and give a response. All test takers must perform all tasks presented to them on the tape; there is no selection of the tasks.

As technology has progressed, new approaches have been developed that rely on various computer technologies, such as computers' adaptive capabilities. The Computerized Oral Proficiency Instrument (COPI) is a computer-adaptive oral proficiency test administered through CD-ROM (Malabonga & Kenyon, 1999). Its design stems from the OPI and the SOPI. The main aim of the COPI is to use computer technology to allow the test taker and computer to work together to produce a speech sample ratable using the ACTFL criteria. The program enables the test taker to perform what he/she can do in a second language. Underlying the COPI is a large pool of speaking assessment item tasks that cover a wide variety of content areas and topics.

The COPI permits test takers to have control of the time they take to prepare for and respond to a task. While a maximum time limit needs to be enforced to ensure the testing process continues, that limit is long enough to allow for most test takers to experience control.

Over the few years, the COPI has become prevailing worldwide. The number of programs/institutions that use computers as the main form of delivery for oral proficiency test has increased dramatically (Garcia Laborda, Magal Royo & Barcena Madera, 2015). However, this situation is not the same everywhere. While countries like the United States or the United Kingdom are spearheads of this tendency, others like Turkey, lag way behind. While the current wave of technology in testing oral proficiency is a bold reminder of Turkey's being left behind in following the speed of change, another sad evaluation has yet to come about Turks. According to the reports of Türkiye Politikaları Araştırma Vakfı (TEPAV) (Koru & Akesson, 2011), the English Proficiency Index (EPI) ranks Turkey 43$^{rd}$ among 44 countries, behind countries such as Saudi Arabia or Indonesia. 44$^{th}$ is Kazakhstan. The survey was investigated by Education First (EF), the world's largest privately held education company. The survey was carried out over the internet and is the most authoritative indicator of English proficiency to date. While it is not a perfect indicator of proficiency, in the absence of a useful cross-country data-set, EPI scores can be used as a proxy to explore cross-country differences. Turkey's poor English proficiency is part of a broader education problem. The high school entrance exam consequently tests students at an extremely rudimentary level on the subject. In some instances, English programs are suffering cuts. Students often took a year of "Hazırlık" or "Preparation" between fifth and sixth grade, during which they received intensive English instruction. This class was put off to be held between eighth and ninth grade in the late 1990s. In 2005, preparation class was called off entirely. Students no longer have a year in which they exposed to English language instruction unless they go on to University, some of which provide such a preparatory year of English. By this time however, students are already 18 years old, and for some of these students, this preparatory programs of universities are the final stage that they are offered with intensive English language course. From this aspect, English language education given in these preparatory schools is highly important in equipping university level students with necessary language skills (Gömleksiz & Özkaya, 2012). In order to provide the most benefits, it seems inevitable for these

programs to assess their program and ongoing processes and, therefore, make every effort to do effective planning, teaching and, of course, testing (Karakuş, 2013). Although the attempt is made to raise the language proficiency level of their students, there may appear problems in each process of planning, teaching and testing. Since the scope of this study is testing of oral proficiency in preparatory schools, the problems related to this will be addressed.

### 1.2. Statement of the Problem

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. 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. However, good pedagogy dictates that if we stress oral skills in teaching, we must also test them. This dissonance between practice and assessment sends an implicit message to Turkish learners of English that this skill is not as valued as the others that are the objects of evaluation. Therefore, due to not given the same evaluative attention as the other skills, they do not experiment with the oral language as much as they do with the written language. This situation, in turn, causes the lack of motivation for the achievement of communicative oral skills in the part of the students. Harlow and 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." 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. As a result, many programs are often left with indirect methods such as paper and pencil tests where students are tested without having their oral proficiency evaluated adequately. On the other side, there are also some preparatory programs in Turkey which prefer to assess oral proficiency through direct testing, that is the face-to-face speaking tests. However, when we have a look at their examinations, we see variations in their understanding of testing oral proficiency. One of the current studies on this topic was conducted by Aydın, Akay, Polat and

Geridönmez (2016) in which they carried out a series of interviews with the administration of 12 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; and 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 with respect to language proficiency tests, the problems they encounter in the process of preparation, administration, and evaluation and their attitudes towards computer assisted language assessment. The most prominent findings of the study showed that (1) the understanding of what a proficiency test should assess differ to a great extent. In some programs, the content of the oral proficiency tests depends on the curriculum and therefore, including the topics covered in the lessons to the proficiency test is highly valued. While these institutions emphasize their being strict in addressing units in the curriculum in their exams, the other programs reject having proficiency tests related directly to classroom lessons. They, on the other hand, indicate testing global language ability of their students in proficiency exams; (2) apart from the difference in the views on what a proficiency test should cover, there appears another difference in the criteria the institutions take as the basis for their proficiency concept and therefore, there is a lack of commonly agreed standards. While most of the institutions state taking The Common European Framework of Reference for Languages (CEFR) as the basis for their proficiency concept, the way they perceive proficiency is highly affected by their context-specific conditions and beliefs. Even if a student receives language education in another institution, the language level they receive is not accepted by most of the schools delivering similar education; (3) all believe the importance of including four skills in a proficiency test; namely reading, listening, writing and speaking. Yet, most of them cannot test speaking skill due to practical reasons; (4) 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; (5) they also state experiencing certain problems in 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 high number of students, tests are provided in multiple-choice format and the statistical analysis of test results are not done by experts in most institutions because of the reasons mentioned above. Within the purpose of this study, particularly in relation to speaking skill, the data gathered from the leading universities of Turkey clearly show that among all skills, testing oral proficiency is referred 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 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.

### 1.3. Significance of the Study

Turkey has 206 universities by the year of 2018 (see www.yok.gov.tr) and nearly 4 million students in total amounts. Almost each university has preparatory programs which aim to enable students to acquire the proficiency and general language skills required for their undergraduate studies. Although the exact number of students in these programs is not known, approximately 55.000 students begin their English language education in these programs each year. Given the pedagogical need to “test what is taught”, preparatory program instructors are faced with the practical dilemma of how to test the oral skills of such a large number of students. In Turkey, to this day, there is no consensus as to a standard administration framework for the testing oral proficiency and this, in turn, brings about several variations and drawbacks in practical and theoretical terms. Therefore, focusing on oral proficiency tests is an issue worth studying because of above mentioned problems and concerns (see section 1.1.). In order to withstand the constraints of time and impracticality, the present research suggests a possible path to develop an oral proficiency test in which the current developments in computer

technologies are presented as a viable alternative for keeping up recent trends, and addressing the demand and potential need for preparatory programs.

This study is significant within the present context of English language testing in Turkey since it brings together conventional language testing practice and the use of new computer technology. It is worth mentioning that the available literature in the field of computers in the assessment of second language ability is very much in favour of the use of technology (Brown, 1997; Alderson, 2000; Chapelle, 2001). Computer-based Tests (CBTs) in general offer many advantages over their paper- and-pencil counterparts. According to Mousavi (2007) for example, there are technical, administrative, and pedagogical advantages associated with testing oral proficiency through CBTs. However, in Turkey, not much has been done in the use of computers in testing speaking. To the researcher's knowledge, there is no study in Turkey with regard to developing and validating a computerized oral proficiency test. This study is also significant due to theoretical and practical concerns. Theoretically, it tries to meet the test usefulness criteria (i.e. reliability, validity and practicality) proposed by Bachman and Palmer (1996) and practically it provides a greater standardization of test administration conditions, and implications for areas such as a) item writing procedures; b) task development; and c) software design. Therefore, this study provides a detailed guidance for other researchers who intend to design and develop a CBT. Therefore, a wide range of professionals in the following areas can benefit from the findings of this study including researchers in test development and task design; researchers in validation and reliability studies in language testing; test administrators, language teaching institutions; English language teachers and learners.

### 1.4. Purpose of the Study

The focus in this study is on the design, construction, administration and validation of the Computerized Oral Proficiency Test of English as a Foreign Language (COPTEFL). Although recent advances in computer technology have promoted 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, 2015a). 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, 2015a; Zhou, 2008). 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 firstly attempted to develop the COPTEFL and then investigate the equivalence of the semi-direct (COPTEFL) and the direct (face-to-face) versions of a test of oral proficiency. The present study is therefore; a 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 COPTEFL measures what it intended to measure. It suggests demonstrating the validity of the test from multiple perspectives. For this purpose, it offers a range of different types of evidence including the processes of test development, test taking and rating. In doing so, it provides deeper insights for a priori test validation specifically investigating theory-based/construct or cognitive validity and context validity for the test task development and a posteriori test validation specifically investigating scoring validity for the test results. Finally, the study also explores test taker attitudinal reactions to taking different formats of oral proficiency. In this respect, it suggests that test taker attitudes might represent an important source to obtain deeper understanding 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. The research questions that guided the present study are as follows:

1. What are the cognitive demands of the speaking test tasks?
2. What are the contextual characteristics of the speaking test tasks?
3. 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?
4. What are the attitudes of test takers in relation to the test delivery modes?

# 2. LITERATURE REVIEW

## 2.1. Introduction

The purpose of this chapter was to review the literature in language testing and technology from which the rationale for the present study was drawn. First, a brief account of the literature on what it means to know a second language was introduced with a focus on theoretical frameworks for language proficiency. In the following section 2.3, speaking ability, the language skill that the current research dealt with, was handled in order to present a brief account for the question of what it means to speak fluently in a second language. Then, validation in language testing and the framework offered by Weir (2005) for validating speaking tests were introduced. Relevant issues regarding validity components i.e. theory-based validity, context validity and scoring validity were also mentioned since they provided validation schemes for both priori and posteriori evidence on the validation of the present test. As the title of the study suggested, this research investigated oral proficiency testing through computer as a delivery-medium. So, in section 2.5, how computer technology contributes to the field of language testing was explained with a particular attention to its use in oral proficiency testing. In the following section, the section 2.6, empirical research on oral proficiency testing was presented. In the last section, the section 2.7, the context of the present study, Turkish context, was examined with regard to the studies in language testing field.

## 2.2. Defining Language Proficiency

In the broadest sense, the term "language proficiency" refers to what an individual can do with the language at the present moment regardless of how s/he learned it. It is the general language ability of an individual. As discussed in Mousavi (2007), one effective way of defining proficiency is to examine what it is not. At this point, it is convenient to distinguish language proficiency from language achievement. While proficiency is general and theory-based, achievement is specific and curriculum-based. In order to get a better understanding, it is perhaps more convenient to contrast proficiency tests with achievement tests. An achievement test is a test which measures how much of a language a learner has learnt in relation to a specific course of study. It is directly related to content of language courses and given as unit, midterm, or final

tests. A proficiency test, on the other hand, is a test which aims to measure how much of a language is learnt. It tries to answer the questions: “Having learned this much, what can the learner do with it?”, “What do they have to do in the language in order to be considered proficient?" The content of a proficiency test, as opposed to achievement tests, is not based on the content of the language courses that learners may have followed.

As stated above, language proficiency is theoretical. The development of a language proficiency test must, therefore, be based on a theory which clearly defines the construct to be measured. One of the most comprehensive frameworks is Canale and Swain (1980) and Canale’s (1983) Communicative Competence model. They proposed four different components of communicative competence: (1) Grammatical competence, (2) sociolinguistic competence, (3) discourse competence, and (4) strategic competence. In this model, *grammatical competence* refers to mastering the linguistic code (knowledge of the rules of grammar and lexis). *Sociolinguistic competence* is concerned with appropriateness of utterances and takes into consideration the appropriateness of what is said, as well as the social context in which the utterance would be deemed appropriate. *Discourse competence* relates to the ability to create unified texts, either spoken or written, in different genres, such as a personal letter, a speech, or a narrative. *Strategic competence* refers to the use of verbal and non-verbal communication strategies that are called into action in order to “compensate for breakdowns in communication due to performance variables or to insufficient competence” (Canale & Swain, 1980, p. 30).

Canale and Swain's (1980) model of communicative competence has been modified over the years. Particularly, Bachman (1990) contributes to it by using the term “communicative language ability”. Bachman preferred to use the phrase Communicative Language Ability (CLA) rather than the more common term, “Communicative Competence” in order to emphasize the fact that he is not only concerned with “competence” but also with “performance”. Bachman's (1990) model is consistent with models of Canale and Swain (1980), and Canale's (1983) later version. The CLA framework includes three components:

(1) Language competence: “A set of specific knowledge components that are utilised in communication via language” (Bachman, 1990, p. 66).

(2) Strategic competence: “The mental capacity for implementing the components of language competence in contextualised communicative language use” (Bachman, 1990, p. 67).

(3) Psychophysiological mechanisms*:* “The neurological and psychological processes involved in the actual execution of language as a physical phenomenon” (Bachman, 1990, p. 67).

Bachman (1990) relates to the interaction of these three components with the language use context and the language user’s general knowledge of the world as in Figure 2.1.

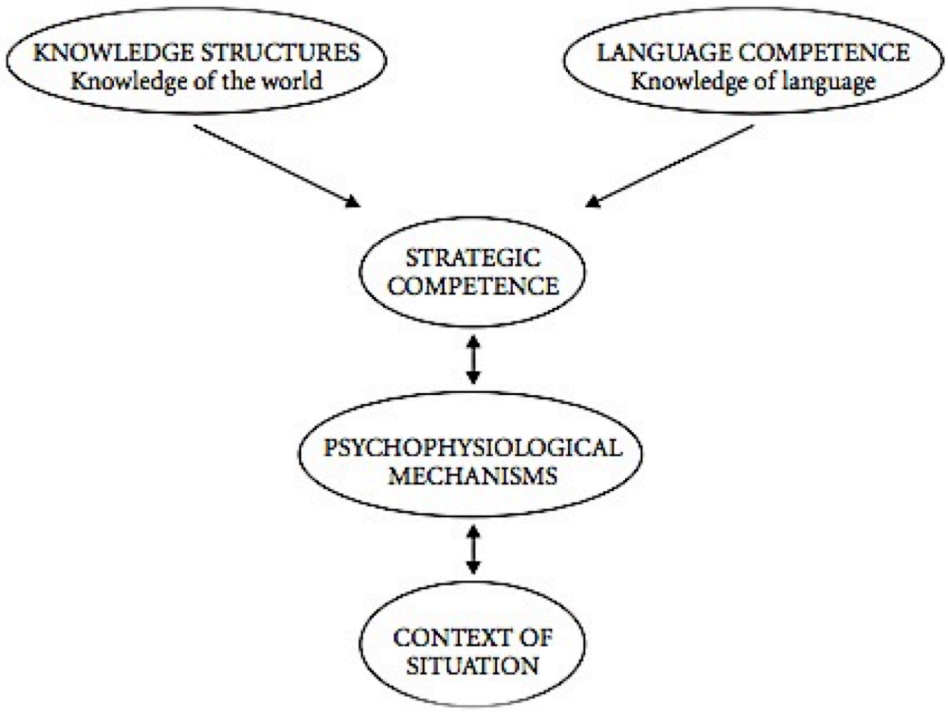



**Figure 2.1.** *Components of communicative language ability in communicative language use (Adapted from Bachman 1990, p. 41)*

Bachman (1990) categorizes language competence into the components in Figure 2.2. He emphasizes that the tree diagram is intended as a visual metaphor which shows the hierarchical relationships among the components. Although they appear as if they are separate and independent of each other, these components all interact with each other and with features of the language use situation (Bachman, 1990).

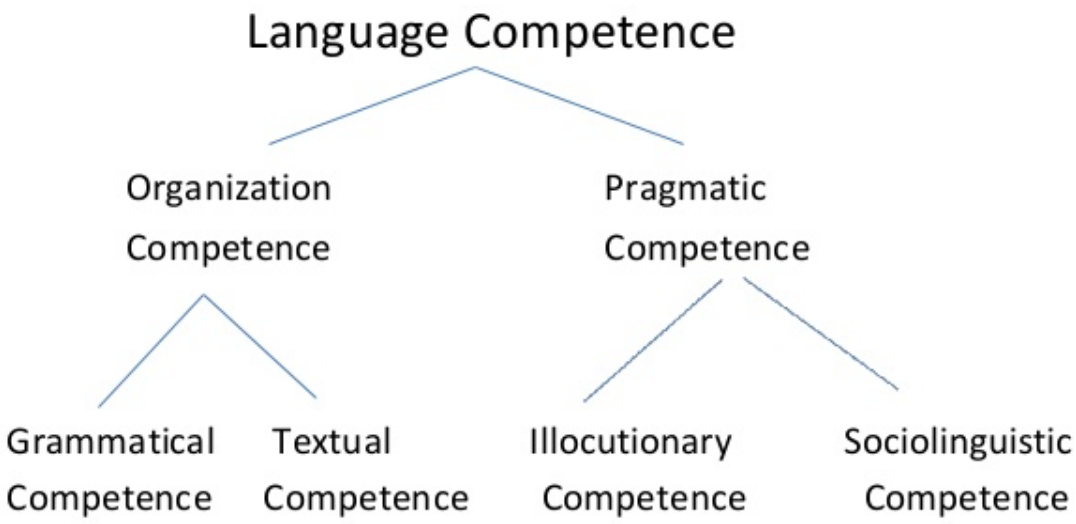



**Figure 2.2.** *Components of language competence (Adapted from Bachman, 1990, p.67)*

A brief description of the main components of Bachman's (1990) model was presented below. Language Competence in Bachman's (1990) model is composed of two components:

(1) Organisational competence, which is divided into two:

a. *Grammatical competence:* Knowledge of vocabulary, morphology, syntax, and phonology/graphology.

b. *Textual competence:*

> "Includes the knowledge of the conventions for joining utterances together to form a text, which is essentially a unit of language — spoken or written — consisting of two or more utterances or sentences that are structured according to rules of cohesion and rhetorical organisation" (Bachman, 1990, p.88).

(2) Pragmatic competence, which deals with the speaker's or writer's ability to achieve his purpose through his utterances. It consists of:

a. *Illocutionary competence:* The ability to express and interpret the function performed in saying something. For example, the statement "It's cold in here" may function as an assertion, a warning, or a request to turn the heater on. The theory of speech acts makes a distinction between an utterance act (just saying something), a propositional act (referring to something), and an illocutionary act. The meaning of

an utterance can thus be described in terms of its propositional content and its illocutionary force.

b. *Sociolinguistic competence:* Sensitivity to, or control of, the conventions of language use that are determined by the features of the specific language use context; it enables us to perform language functions in ways that are appropriate to that context.

Bachman's goal in identifying the various types of competencies described here is to construct adequate tests of second language learners' proficiency; his re-organising and re-defining of the terms used previously by Canale and Swain (1980) is motivated by a desire to make the terms more testable.

Even though Bachman's CLA framework is not very different from earlier frameworks of communicative competence, it seems

> "innovative in its attempt to bring earlier components of communicative competence together into a coherent whole and represent the processes by which the various components interact with each other and also with the context in which language use occurs" (Joo, 2008, p.56).

In this section, a brief discussion on language proficiency models was presented. Since this study focused on speaking ability as a language proficiency construct, it was necessary to define what it means to speak fluently in a second language, types of tasks that can be used in testing this ability and what criterial levels are to be used while scoring the speaking performances.

### 2.3. Defining the Speaking Ability

The ability to speak a second language is a subset of a learner's overall language ability —or proficiency—in the target language. So, the question of what speaking ability is, is closely related to the question of what it means to know a second language (for *language proficiency* see section 2.2).

In order to define what it means to speak fluently in a second language, we should have an understanding of the elements necessary for spoken production. These elements are the properties of spoken language that are presupposed for the ability to speak fluently. They involve the knowledge of language features and the ability to process information and language "on the spot" (Harmer, 2001).

Speakers' productive ability involves the knowledge of language features including (Harmer, 2001): (1) Connected speech; an effective speaker is able to produce fluent connected speech in which the sounds are modified (assimilation), omitted (elision), added (linking r) or weakened (through contractions and stress patterning), (2) expressive devices; in order to be effective communicators, learners should be able to vary the pitch and stress of particular parts of utterances, change volume and speed to show how they are feeling, (3) lexis and grammar; this is the knowledge of lexical phrases and their functions such as agreeing or disagreeing, expressing shock or approval, and (4) negotiation language; it refers to the negotiatory language that is used to seek clarification (i.e. 'I am sorry, I don't understand' or 'Could you explain that, again?' ) and to show the structure of what we are saying (i.e. 'What I mean is' or 'To begin with/And finally').

Speakers' success is also dependent on the ability in rapid mental/social processing that speaking necessitates. This includes (Harmer, 2001):

(1) Language processing; which involves putting language into coherent order in order to convey the meanings that are intended,
(2) Interacting with others; this involves interaction with one or more participants, a great deal of listening as well as understanding of how participants feel, and the knowledge of how to take a turn or allow the others to do so, and
(3) (On the spot) information processing; speakers need to be able to process the information they are told the moment participants provide.

For the purpose of testing speaking ability, it is crucial to set tasks that elicit behaviour which truly represents the test takers' ability and which can be scored validly and reliably. In order to utilize the notion of tasks in the design and development of language tests, it is necessary to describe what it is meant by this term. The following section presented a brief account for the term 'task' and speaking task types (see section 2.3.1).

### 2.3.1. Types of speaking tasks

Generally speaking, tasks refer to any activity in which students are engaged with comprehending, manipulating, producing or interacting in a target language with a focus on meaning (Nunan, 1993). In terms of the definition of a speaking task, it is that given by Luoma (2004, p. 31): "Activities that involve speakers in using language for

the purpose of achieving a particular goal or objective in a particular speaking situation." Luoma (2004) divides speaking tasks into two major groups: open-ended and structured tasks. Each has a set of different task types as given in the Table 2.1 below:

**Table 2.1.** *Speaking tasks (Luoma, 2004, p.48-50)*

| **Task Types** | | |
|---|---|---|
| **Open-ended Speaking Tasks** | | |
| | **Discourse type tasks** | **Role play tasks** |
| | Descriptive | Social simulation |
| | Narrative | Professional simulation |
| | Instruction | |
| | Comparison | |
| | Explanation | |
| | Justification | |
| | Prediction | |
| | Decision | |
| **Structured Speaking Tasks** | | |
| | Reading aloud | |
| | Sentence repetition | |
| | Sentence completion | |
| | Factual short-answer questions | |
| | Reacting to phrases | |

The open-ended tasks are the ones which require learners to do something with the language as an indication of their speaking ability. They consist of discourse type and role play tasks. Tasks performed in a speaking test in which the test taker is expected to generate connected discourse are considered to be instances of discourse type and role play tasks. They are relatively long activities such as giving a presentation, or a short, function based action like making a request. One way of dividing open-ended speaking tasks into task types is discourse types, which include *description, narrative, instruction, comparison, explanation, justification, prediction, and decision tasks*. Another category of open-ended tasks is role-play. Some of these tasks may simulate the professional context of the test takers and put the testers in the role of others such as their clients or non-expert acquaintances. Some of the others may simulate the social or service situations such as buying something or going to a restaurant. The structured speaking tasks are, on the other hand, "the speaking equivalent of multiple choice tasks and therefore the expected answers are usually short, and the items tend to focus on one narrow aspect of speaking at a time" (Luoma, 2004,

p. 50). These include the types of tasks as *reading aloud, sentence repetition, sentence completion, factual short answer questions and reacting to phrases*.

### 2.3.2. Criterial levels of speaking performance

The qualities of speaking ability cannot be measured directly as in the measurement of height or weight. In order to score test takers' speaking ability validly and reliably, rating scales against which test takers' performances are judged should be provided for the specifications of the skill. Examples of these are American Council on the Teaching of Foreign Language (ACTFL) Guidelines, the Interagency Language Roundtable (ILR) and the Global Scale of English (GSE). A description of 'Intermediate' ability in GSE is provided as an example below (see Table 2.2):

**Table 2.2.** *Global Scale of English: Speaking assessment descriptors. (Adapted from Global Scale of English Assessment Framework, 2016, p.8)*

| Speaking Assessment Descriptors | GSE 51–58/B1+ |
|---|---|
| **Spoken production and fluency**<br>Length of contribution<br>Intelligibility<br>Pausing & hesitation<br>Cohesion<br>Repair strategies | Can communicate using longer stretches of connected clauses and functional language (e.g. compare/contrast; reason/explanation). May pause when handling more complex matters and need to use repair strategies. Is generally intelligible and can use basic stress and intonation to support meaning. |
| **Spoken interaction**<br>Ability to maintain or develop interaction<br>Coherence<br>Appropriacy<br>Pragmatic Strategies | Can maintain and develop the interaction with relative ease and coherence on familiar topics, including some abstract matters. Can respond with some flexibility within familiar topic areas and can reformulate simple responses but may need to ask for clarification in less familiar topic areas. |
| **Range**<br>Structure<br>Vocabulary<br>Discourse/communicative functions | Can maintain and develop the interaction with relative ease and coherence on familiar topics, including some abstract matters. Can respond with some flexibility within familiar topic areas and can reformulate simple responses but may need to ask for clarification in less familiar topic areas. |
| **Accuracy**<br>Structure<br>Vocabulary<br>Appropriacy of vocabulary<br>Functions | Uses a range of words, structures and collocations. Can use functional language to deal with unfamiliar everyday topics but has a limited range of complex language. Searches for unknown words and /or uses repetitive paraphrases. |

Whatever rating scale ACTFL, ILR or GSE is adopted, the important thing is for tester to be as clear as possible (before the test is developed) about the nature of the ability that is to be tested (Hughes, 1992). Scoring will be valid and reliable only if (Hughes, 1992):

1. Clearly recognizable and appropriate descriptions of criterial levels are written and scorers are trained to use them.
2. Irrelevant features of performance are ignored.
3. There is more than one scorer for each performance.

In each step of test development, the goal is to provide evidence in relation to the two essential aspects of test development, i.e. validity and reliability. Accordingly, since this is a test development study, the issues of validity and reliability (or scoring validity) were discussed thoroughly in the following sections and a test validation model that forms the theoretical basis of the test validation process in the present study was presented.

### 2.4. Validation in Language Testing

The theoretical basis of language testing is heavily influenced by the theories of validity developed in the general measurement literature (McNamara, 2010). With more than a hundred years of development, validity theory has evolved significantly and how validity is conceived and presented has varied over the years.

The standard definition of validity in the first half of the twentieth century was that it was the degree to which a test or examination "measures what it purports to measure" (Ruch, 1924). With this traditional definition, the key validity question has always been "Does my test measure what I think it does?" (Fulcher, 2010, p.19). The answer was mainly judged with regard to how well a test predicted the criterion that it was used to predict. The early focus of validation, therefore, was on the prediction of specific criteria, which later validity theory was called criterion-related validity (Luoma, 2001).

Originally, theorists divided validity into four separate categories; content, concurrent, predictive, and construct validity. The concurrent and predictive categories were subsequently identified to constitute subcategories of a single type, i.e. criterion-related validity (Luoma, 2001). Relevant evidence of this segmented approach to validation can be found in several language assessment research studies reported in

journals and textbook chapters of the period (Kunnan, 2013). That practice was based on an earlier belief that the validity is always specific to the purpose of the test and different kinds of test require different evidences of validity. Although the division of validity into types unified practice, there was also increasing dissatisfaction with its shortcomings. Luoma (2001) explicated the problem in a nutshell: "By the 1970s, measurement theorists began to feel unhappy with the mechanical and simplified way in which commercial testing boards operationalized it [validity]" (p.68). What bothered theorists is reflected in Anastasi's (1986) criticism:

> "Thus test constructors would feel obliged to tick [the three types of validity] off in a checklist fashion. It was felt that they should be covered somehow in three properly labelled validity sections in the technical manual, regardless of the nature or purpose of the particular test. Once this tripartite coverage was accomplished, there was the relaxed feeling that validation requirements had been met" (p.2).

Eventually, treating these as separate types of validity was considered unsatisfactory, and rejected in favor of a unitary view of validity as being entirely construct validity (Carr, 2011). Thus, the former validity "types" turned out to be *types of evidence* supporting the construct validity of a test –that is "supporting the argument that score-based inferences and interpretations regarding examinee ability are appropriate" (Carr, 2011, p.153). As early as 1988, Messick argued that relying entirely on one type of validity evidence is not appropriate and thereafter, he pointed that each type of validity is "a complementary form of evidence to be integrated into an overall judgment of construct validity" (Messick, 1988, p. 37). He argued that each separate type of validity is not enough by itself and each of them is "a complementary form of evidence to be integrated into an overall judgment of construct validity" (Messick, 1988, p. 37). In his unified concept of validity, Messick (1989) integrated considerations of content, criteria and consequences into a construct framework for the empirical testing of hypothesis about score meaning and noted score meaning and social values in test interpretation and test use. Messick (1989) defined validity as "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (p. 13). Messick's influential definition implies that to be able to justify that a specific score interpretation is meaningful and appropriate, supporting evidence is required. To gather such evidence, one first needs to

define the construct to be measured. Construct validity is therefore about the extent to which a particular interpretation is justified in relation to the construct measured.

From the earlier and current definitions of validity, it can be understood that there is a shift of focus from the validity of the test itself to the validity of the test scores, and the interpretations and inferences made based on these scores. With the introduction of the unitary notion of validity, measurement specialists began to consider construct validity to be the central validity concern because validation is no longer regarded as mere quantification of the predictive power of a test and because the current view is that the interpretations necessarily involve constructs. Hence, construct validity is about any evidence regarding the interpretations of score meaning and all sources of validity evidence are actually a part of construct validity (Messick, 1988).

One widely acknowledged model for considering the full manifestations of conceptions for construct validity is proposed by Messick (1988; 1989). Messick saw assessment as a process of evidence collecting in order to make inferences about individuals and saw the task of establishing the meaningfulness and defensibility of those inferences as being the chief task of assessment development and research. Messick, from a different point of view, saw tests as instruments which are used in society, and his argument is that the meanings of test scores cannot and should not be inquired without reference to the way they are going to be used. Therefore, according to him, inquiry of validity should always entail investigation into the values and consequences involved in the interpretation and use of test scores. In order to make his conception of validity more comprehensible, he proposed a new model for the concept. This was the first time concepts such as value implications and social consequences were involved within the model of assessment validation. Messick (1989) termed his faceted conception of unified validity "the progressive matrix" (see Table 2.3).

**Table 2.3.** *Facets of validity (Messick, 1989, p.20)*

| | **Test Interpretation** | **Test Use** |
|---|---|---|
| **Evidential Basis** | Construct validity | Construct validity + Relevance / utility |
| **Consequential Basis** | Value implications | Social consequences |

When Messick's framework is read as a progressive matrix with the different facets contributing to this unified validity concept, the overall influence of construct

validity and the critical importance of each facet become clearer. It may be useful to gloss it rather abstract terms in the following interpretation of the matrix:

**Table 2.4.** *Understanding Messick's validity matrix (Adopted from McNamara & Roever, 2006, p.14)*

| | **What test scores are assumed to mean** | **When tests are actually used** |
|---|---|---|
| **Using evidence in support of claims: test fairness** | What reasoning and empirical evidence support the claims we wish to make about candidates based on their test performance? | Are these interpretations meaningful, useful and fair in particular settings in which the test is to be used? |
| **The overt social context of testing** | What social and cultural values and assumptions underlie test constructs and hence the sense we make of scores? | What happens in our education systems and the larger social setting when we use tests? |

This progressive matrix view of Messick's fourfold classification of facets of validity. In this framework, Messick brings together the two aspects of validation -the quality of the inferences about abilities that are represented by test scores, and the need to validate the relevance and usefulness of tests in particular contexts – with two new features: a concern for the values implicit in test constructs, and the social and educational consequences of their use.

Since the testing era has shifted from the period that focused on the division of validity to construct validity as the main concern, validity has become a complex concept (Luoma, 2001). Messick's conception of validity made validation seem unapproachably complex for some (Shepard, 1993). However, many researchers found the complexity appropriately captured the challenges inherent in validation (Chapelle, 2013). Implications of Messick's framework specifically for language testing was outlined by Bachman and Palmer (1996). They built on Messick's concepts of test validation by articulating a theory of test usefulness, which they see as the most important criterion by which tests should be judged. In doing so, they explicitly incorporate Messick's perspective for construct validity, but also add dimensions that affect test development in the real world. Test usefulness, in their view, consists of six major components or what they call "critical qualities" of language tests, namely construct validity, reliability, impact, interactiveness, authenticity and practicality. Section 2.4.1 below presented the test usefulness model by Bachman and Palmer (1996).

### 2.4.1. Test usefulness

The most important consideration in devising a language test is the use for which it is intended, so that the most important quality of a test is its usefulness (Bachman & Palmer, 1996). As stated in Bachman and Palmer (1996), simply using a test does not make it useful. In their book, they suggest a model of test usefulness which guides for quality control throughout test development process. This model includes six test qualities as (1) reliability, (2) construct validity, (3) authenticity, (4) interactiveness, (5) impact, and (6) practicality (see Figure 2.3).

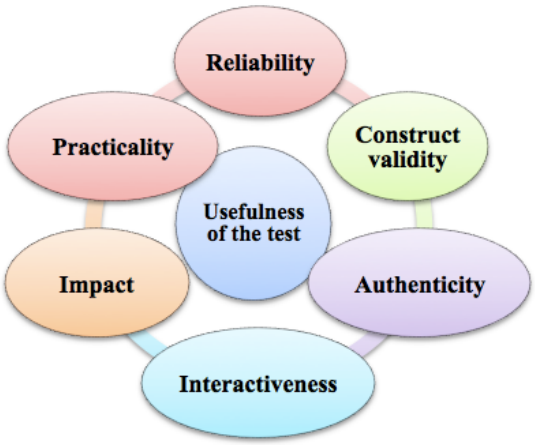



**Figure 2.3.** *Qualities of language tests (Bachman & Palmer, 1996, p.17)*

(1) Reliability*:* Reliability refers to consistency of measurement. A reliable test score is the one that is consistent across different characteristics of testing situation. This can be illustrated as in Figure 2.4.

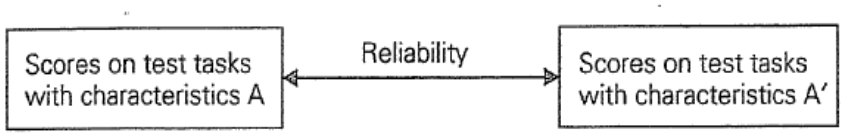



**Figure 2.4.** *Reliability (Bachman & Palmer, 1996, p.20)*

(2) Construct validity: In their model, Bachman and Palmer (1996, p. 21) refer to construct validity as "the meaningfulness and appropriateness of the interpretations" that are made on the basis of test scores. In other words, for test scores to be meaningful researchers need to be able to present evidence that the scores reflect those areas of language ability which are intended to be measured.

(3) Authenticity: It pertains to the degree of correspondence between performance on the test and the language use in specific domains other than the test itself. It is the extent to which the tasks required on a given test are similar to normal real-life language use.

(4) Interactiveness: Bachman and Palmer (1996) refer to interactiveness as "the extent and type of involvement of the test-taker's individual characteristics in accomplishing a test task" (p.25). The individual characteristics that are most relevant for language testing are reported as the test taker's language ability, topical knowledge, metacognitive strategies, and affective schemata. These individual factors can thus characterize the nature and the extent of interactiveness between the test and the individual taking test.

(5) Impact: As another test usefulness quality, impact refers to the prospective and potential behavioral changes that may occur in education and society as a result of the introduction of a new test or testing approach. Washback or backwash is addressed by Bachman and Palmer (1996) as "an aspect of impact that has been of particular interest to both language testing researchers and practitioners" (p.30). It is generally conceptualized as a continuum, ranging from positive to negative.

(6) Practicality: Bachman and Palmer (1996) address the different nature of practicality when compared to the other five qualities. While the other five qualities base on the *uses* that are made of test scores, practicality pertains to the ways in which the test will be implemented, and, to a large degree, whether it will be developed and used at all. That is, if the resources needed for implementing the test exceed the resources available, the test will be impractical, and will not be employed if resources cannot be allocated more efficiently, or if additional resources cannot be allocated as well. So, the question, at this point, is how to test practicality. According to Bachman and Palmer (1996), in order to assess practicality we need to specify what we mean by resources. They classify resources into three general types as (1) human resources (i.e. test writers, scorers, or raters), (2) material resources (e.g. space –rooms for test development and test administration, equipment –tape and video recorders, computers, and materials – paper, pictures), and (3) time (i.e. development time, time for specific tasks) (see Figure 2.5).

1 **Human resources**

(e.g. test writers, scorers or raters, test administrators, and clerical support)

2 **Material resources**

*Space* (e.g. rooms for test development and test administration)
*Equipment* (e.g. typewriters, word processors, tape and video recorders, computers)
*Materials* (e.g. paper, pictures, library resources)

3 **Time**

*Development time* (time from the beginning of the test development process to the reporting of scores from the first operational administration)

*Time for specific tasks* (e.g. designing, writing, administering, scoring, analyzing)

**Figure 2.5.** *Type of resources (Bachman & Palmer, 1996, p.37)*

As mentioned above, the notion of test usefulness is described with regard to six qualities. It cannot be, however, offered general prescriptions for a test about what the plausible balance among qualities should be. Two of the qualities –validity and reliability, on the other hand, are crucial for tests, and are sometimes addressed as *essential measurement* qualities (p.19). The reason for this is that these two are the qualities that provide the main justification for using test results as a basis for making decisions. What is less clear in the Bachman and Palmer's (1996) account of test usefulness is how these various qualities should be measured and weighted in relation to each other. At this point, a useful framework that can guide a validation study in its necessary steps is the socio-cognitive framework proposed by Weir (2005). In this study, the guidance was taken from the socio-cognitive framework in attempts to justify the validity of the speaking test tasks developed to test general speaking skills of learners of English as a Foreign language (EFL).

### 2.4.2. Weir's socio-cognitive validity framework

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

The most useful starting point for the test development is to have a framework of validation on which the study might rely to support the systematic gathering of validity evidence to support the claims made for the tests. That is, 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. As stated in the previous part, 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 a 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, see Figure 2.6) was used as the major reference by which the speaking tasks of the study were developed.

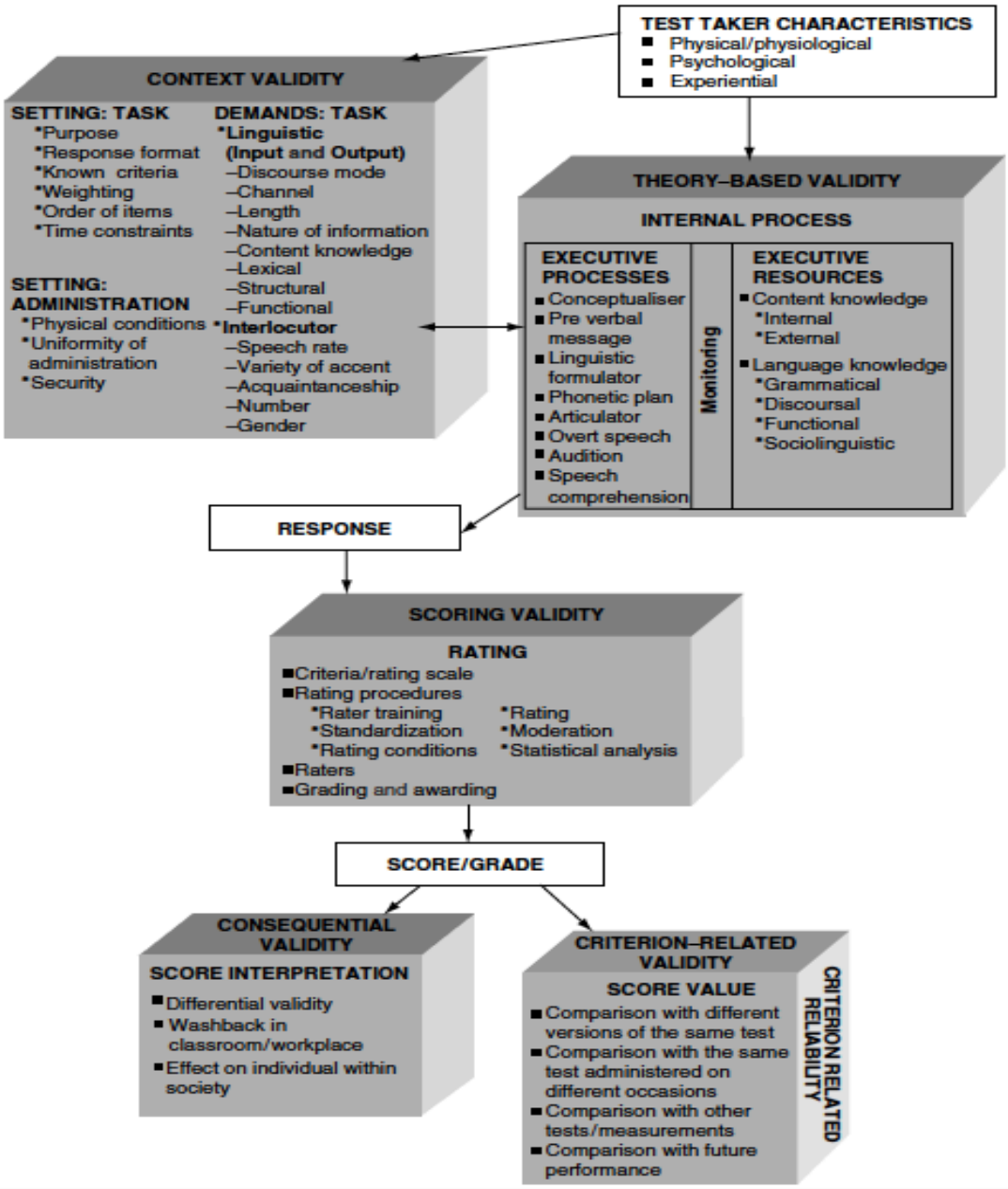



**Figure 2.6.** *A Socio-cognitive framework for validating speaking tests (Weir, 2005, p.46)*

The framework offers 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 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 measurement that assess the same ability. Finally, consequential validity refers to the impact of tests and test scores interpretations on teaching, learning, individuals and society.

As noted in Chapter 1, 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 that included in the study were discussed in the following part of the study.

#### 2.4.2.1.1. *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. It is concerned with the internal mental processes and includes executive resources and executive process. Executive resources involve language knowledge (grammatical knowledge: lexical and syntax), textual knowledge, functional (pragmatic) knowledge, and sociolinguistics and content knowledge of the speaker. These are equivalent to Bachman's (1990) language competence components of organizational and pragmatic competence. Executive processing involves goal setting, monitoring, visual recognition, and pattern synthesizer. 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 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 the dialogue or looking at pictures will influence how the test taker conceptualizes and processes these messages as pre-verbal message (Zainal Abidin, 2006).

In addition to Weir's theory-based validity, the speaking skill descriptors provided by the GSE, 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.

***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. With the current view, at the heart of the validity is the meaningfulness and precision of test scores in relation to the measurement of an

underlying construct. Since from this perspective, test scores "represent the most visible and tangible evidence of the outcomes of an appropriate assessment, it is not surprising that performance outcomes have been relied upon for many years to help determine the validity and reliability of a given assessment" (East, 2016, p.10).

Such an approach needs to examine how test users can know what a test score means and what it can be used for. For example, if a test taker is given a score of 75, B2 or NOVICE HIGH, how can test users know what the score means and how it can be used? (Chapelle, 2013). Test scores tell us something meaningful in relation to the quality we aim to measure (East, 2016). 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). Test scores today tend to be predicted in terms of real-world abilities. As stated in Chapelle (2013), in current approaches, scores are interpreted with regard to pre-determined standards of knowledge and such standards are "typically couched in practical, functionalist terms, reflecting the functionalist, communicative tradition of language teaching" (p.8). For example, the increasingly used Common European Framework of Reference for languages (CEFR, Council of Europe, 2001) represents an ordered set of statements about aspects of communicative ability. The framework described in CEFR shows substantial similarities with Bachman and Palmer's (1996) model of CLA, but it incorporates linguistic processing as different from what is proposed in CLA. The CEFR makes explicit references to linguistic processing skills:

> "Successful task accomplishment may be facilitated by the prior activation of the learner's competences, for example, in the initial problem-posing or goal setting phase of a task by providing or raising awareness of necessary linguistic elements, by drawing on prior knowledge and experience to activate appropriate schemata, and by encouraging task planning or rehearsal. In this way the processing load during task execution and monitoring is reduced and the learner's attention is freer to deal with any unexpected content and/or form-related problems that may arise, thereby increasing the likelihood of successful task completion in both quantitative and qualitative terms" (Council of Europe, 2001, p. 158-159).

The CEFR was developed by the Council of Europe (2001) with the aim of describing what language learners need to learn to be able to communicate through a language and what knowledge and skills they need to develop to be able to do so. Although it was developed in an attempt to provide a common ground for the development of syllabuses for language courses, curriculum guidelines, examinations, course materials or textbooks across Europe (Council of Europe, 2001), its influence has expanded beyond Europe and today, many test developers feel the need to align their tests to the CEFR even though its authors note that the CEFR is not a how –to-

guide for language test construction. According to the complexity of skills concerned, the CEFR provides a set of six common reference levels that describe language proficiency ranging from lowest to highest: A1, A2, B1, B2, C1, C2. 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 travelling 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).

As stated in Chapelle (2013), the wording of these descriptors is important because they represent the test *construct*, "the assumed view of language proficiency which is assessed in the test and which is the target of measurement in any individual case" (p.9). 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 Global Scale of English (GSE). Although the GSE proficiency scale was created with reference to the CEFR, the main difference between the GSE proficiency scale and the CEFR proficiency scale stems from its granular structure (see Figure 2.7).

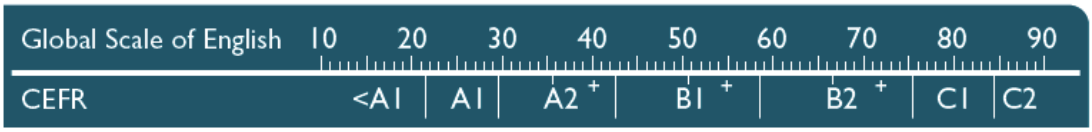



**Figure 2.7.** *Global Scale of English aligned with the CEFR*

As shown in Figure 2.7, the GSE is a proficiency scale from 10 to 90, which is aligned to the CEFR, presenting a more granular measurement of proficiency within a single CEFR level (GSE, 2015). The GSE 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".

The GSE descriptors for the overall speaking ability are shown in Appendix A. The fact that the GSE identifies language proficiency in different levels and offers illustrative descriptors of "can-do" activities at each range of a proficiency level makes it useful reference in task design especially when a specific range is targeted for a task. These descriptors are used in the study to guide for the alignment of the tasks to the different proficiency levels.

#### 2.4.2.1.2. *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. 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 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. For instance, "if it can be shown that the response format such as a short presentation followed by a discussion, reflects the instructional objectives of a course on speaking for academic purpose, then this is one piece of evidence for the choice" (Zainal Abidin, 2006, p.64).

#### *2.4.2.1.3. Scoring validity*

Scoring validity is concerned with all test aspects that can influence scores' reliability. Weir (2005) defines it "the degree to which examination marks are free from

errors of measurement and therefore the extent to which they can be depended on for making decisions about the candidate" (p.23). 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, or unreliability. As it was put forward in Geranpayeh (2013) problems of inconsistency or test administration 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 (see Figure 1). 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).

The choice of test delivery medium -computerized or face-to-face has fundamental implications for the aspects of a test's validity, including the construct, the task characteristics and cognitive demands triggered by the tasks (Galaczi, 2010). Since this study is about the development of a computerized oral proficiency test, a discussion of using computers in testing speaking is within the scope of the study. There are several test-delivery formats for oral proficiency tests. These include such formats as direct testing (i.e. face-to-face interviews) and semi-direct testing (tape-based tests or computer-based tests). Semi-direct tests are designed to be a surrogate for the oral proficiency tests in situations where a face-to-face interview is not possible or desirable. 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 with a variety of multi-media input sources (Zhou, 2015). Resulting from adopting computer technology in delivering and administering oral proficiency tests, there have been renewed debates over the appropriateness of two different testing formats, namely face-to-face, or direct testing, and person-to-machine, or semi-direct testing. In order to enhance test efficiency by embracing technological advancements and provide alternative ways to be a surrogate for oral proficiency test in cases where a face-to-face oral proficiency testing is not practical or desirable, it was intended in this study to use

computers as a test-delivery format for oral proficiency tests. The following chapter reviewed the use of computers in language testing.

## 2.5. Computers in Language Testing

The widespread accessibility to large, networked computer labs at education programs and commercial testing centers help increase the application of computer technology in the field of language testing. Computer-based Testing (CBT) –the use of computers to deliver, score, select items, and report scores of assessments has promoted the reliability of the test resulting from the standardized of delivery, the efficiency of test administration, the speed of score reporting and the flexibility in the presentation of tasks with several sources of multi-media input (Zhou, 2015). Computer-based tests have, therefore, received a large amount of attention in the field of language testing, and are now being used as viable alternative and a complement to the more traditional paper-and-pen tests on numerous assessment fronts from high-stake commercialized tests (i.e. Internet-based Test of English as a Foreign Language –TOEFL IBT) to low-stakes available for free on the Internet. There are, of course, some drawbacks in such tests.

It is argued that CBT is limited in the item types they allow. The multiple-choice, cloze and gap-filling techniques are frequently used where other techniques might be more appropriate, but are relatively more difficult to implement in a setting where responses must be machine-scorable (Alderson, 2000). Another drawback is the security of question bank and test administration process. Even though this is also a case for paper-and-pen tests, computer-stores test items and test data are vulnerable to cyber-attacks. In order to ensure the security, a special software might be needed and this may increase the cost of testing. Computer literacy can be a disadvantage, as well. Reading a text from a screen or writing and speaking to a computer screen may be a cause of discomfort on the part of the test takers, which may, as a result, affect their real performance.

Computer-based tests are currently delivered in various forms (Ockey, 2009). Software versions of computer-based tests are delivered on CD-ROMs, however, these are being replaced by Web-based Testing (WBT), which is an area of rapid growth for CBT over the past decade.

### 2.5.1. Web-based testing (WBT)

WBT shares many common characteristics with CBT, but using the Web as its delivery medium. It is delivered via the World Wide Web (WWW), and written in the language of web, HTML (Hypertext Markup Language). The test consists of one or many HTML file(s) located on the tester's computer, the server, and can be downloaded to the candidate's computer, the client. The process of downloading can occur for the entire test at once, or item by item. There are different kinds of WBTs that change due to their technological sophistication. Depending on the need and the budget, test developers could use low-tech or high-tech WBT. In a low-tech WBT, tests run completely clientside and the server is used only for retrieving items and scoring responses (Roever, 2001). This type of WBT is the easiest to build and maintain since the tester does not engage in serverside programming. If item pool is small and adaptivity for the selection of the next item is unnecessary, low-tech WBT is preferable. A high-tech WBT, however, makes heavy use of server and requires adaptive algorithms for the item selection. If item pool is large, complex adaptive algorithms are necessary and the budget allows for the purchase of expensive software and the hiring of computer experts, a high-tech WBT is preferable (Roever, 2001).

Carr (2006) and Ockey (2009) refer to several advantages of WBT, including the following examples: (1) Storing responses on a server can promote the security of the system since it negates the need to store the test on many small-scale servers or computers, (2) candidates who have used the Internet are more likely to be familiar with the WBT interface than with computer programs developed for specific assessment instruments. This mostly disarms the concern that test scores will reveal invalid results due to test takers' lack of computer familiarity with computers. But, still this concern should not be dismissed as an irrelevant factor in affecting the validity of scores, and (3) costs may be limited because tests can be developed to function with browsers that are already available.

#### 2.5.1.1. *Validating web-based tests*

Validation of web-based tests is not different in principle from validation of other tests. But, there are specific validity issues because of the testing delivery format. These issues are briefly explained in Roever (2001). Accordingly, in order to ensure the validity of scores, the following concerns should be taken into consideration:

(1) Computer familiarity: Test takers' varying familiarity with computers can affect their scores and cause construct-irrelevant variance. Tutorials and standard web-browsers, which are more likely to be acquainted with can be used to eliminate this affect. For example, Roever (2001) revealed no significant correlation between self-assessment scores of Web browser familiarity and scores on a Web-based test of second language pragmatics taken by 61 intermediate-level English as a Second Language (ESL) learners.

(2) Typing speed: Test takers' varying typing speed is potentially more serious sources of error variance and is not amenable to quick training. Increase in item response time may minimize the influence of this factor.

(3) Delivery failures and speediness: In the design of web-based tests, it should be ensured that the test does not skip items in delivery process due to technical problems.

(4) Loading time and timer: If the test is delivered through Web, not the clientside, download times can be negligible or considerable due to some factors such as server traffic, complexity of the page, client computer speed, and a host of other factors. So, it is important for timed tests that the timer stops during downloads and restart when the screen is fully displayed.

### 2.5.2. Computer-adaptive testing (CAT)

The most referred advantage of CBT in the literature is the use of computer-adaptivity (i.e. Chalhoub-Deville and Deville, 1999). Computer-adaptive testing (CAT) is a variation of tailored testing which roughly estimates the test taker's ability and select the next test item based on the test taker's estimated level of ability (Aldreson, 2000). CAT is depended on a bank of items, all calibrated on a single-difficulty scale. The difficulty of items is ranged against this scale and preprogrammed into the CAT. Item Response Theory (IRT) allows the tool by which this can be done (Kunnan, 1998). By matching the difficulty of the test item as closely as possible with the ability of the candidates, it is claimed that (Chalhoub-Deville and Deville, 1999) computer-adaptive tests are more efficient measures of ability than conventional paper-and-pen tests since CAT selects a subset of items corresponds to each test taker's ability level.

Consequently, CAT requires fewer items in tests to estimate test takers' abilities with the same degree of precision as linear tests.

A practical advantage of CAT is that the CAT algorithm enhances test security because each candidate is administered a different set of test items depending on his/her language ability level and so, there occurs minimal chance of copying from one other. There is also minimal chance for candidates to encounter the same test items on successive administrations, thus CAT permits the test to be used over and over again on the same test takers. However, CAT requires test developers to create a large number of test items and also requires a large number of test takers for item piloting and calibration, as well. CAT development is costly and demanding, which requires a high level of expertise in computer technology and psychometrics as Dunkel (1997) addressed: "It takes expertise, time, money, and persistence to launch and sustain a CAT development project. Above all it takes a lot of team work" (p. 3-4).

Current CATs are unable to include to extended response types of items such as interviews because the responses cannot be scored on-line and therefore, tasks cannot be calibrated and selected immediately based on the test takers' performances on previous items. While test takers' performances on computerized tests can be collected, human scorer is still required to judge and rate such performances.

One of the advances brought by CBT for testing oral abilities is automated scoring. With the innovations in new technology, automated scoring is becoming popular in language testing environments. Although automated language scoring programs are being implemented in a number of tests (i.e. the Versant series of speaking, in which test takers' responses are rated by speech processing technology), the range of expected responses is very limited, making the job of the automated scoring program easier (Douglas, 2014). At the current point in the development of automated scoring, we should probably conclude that some degree of scepticism is warranted and addresses computer scoring of speech in low-stakes or very highly controlled conditions (Douglas, 2014). As pointed out in Norris (2001): "It is doubtful that the complexities of [speaking] performances and the inferences that we make about them will be captured by automated scoring" (p. 99).

### 2.5.3. Computers and oral proficiency testing

The use of computer for testing speaking ability has recently become more

widespread and there is no doubt that it will become even more predominant in the future. Generally speaking, computer-based tests are mostly devised and implemented for high-stake testing purposes because of the necessity to provide oral proficiency tests for large group of test takers which can be delivered quickly and efficiently while maintaining high-quality. The efficiency of computer-based tests led to the development of various computer-based spoken language tests such as the Computerized Oral Proficiency Instrument (COPI), Basic Skills English Test (BEST) Plus, Digital Video Oral Communications Instrument (d-VOCI), English Speaking Proficiency Test (ESPT), Purdue's Oral English Proficiency Test (OEPT), and SAM Interview (for oral proficiency test formats see Table 2.5).

**Table 2.5.** *Test-delivery formats for speaking tests in foreign/second languages (Yu, 2006, p.27)*

| **Paper and Pencil Indirect** | **Face-to-Face Direct** | **Audio-recorder & Booklet Semi-direct** | **Computer; Online Semi-direct** |
|---|---|---|---|
| Multiple choice tests | OPI | SOPI | COPI, BEST, DVOCI, ESPT, OEPT, SAM Interview |
| Written Responses | | | |

The COPI, Computerized Oral Proficiency Instrument created by Malabonga and Kenyon in 1999 at the Centre for Applied Linguistics (CAL) in the United States is perhaps the first computer-mediated test of speaking ability. CAL develops COPI as an alternative to the Oral Proficiency Interview (OPI) and the Simulated Oral Proficiency Interview (SOPI) (see section 1.1). COPI is based on a computer-adaptive algorithm that test takers' choices of levels during the practice task determine their level of difficulty for the first task. After the first task, however, the algorithm alternates from giving test takers choices or not. During the times that test takers are not provided with a choice, the algorithm pushes test takers to higher level tasks to ensure that they are given an opportunity to reach their ceilings.

Kenyon and Malabonga (2001) discuss that the COPI can be considered as an advanced version of the SOPI, but the COPI has not been operationalized to test oral proficiency because it was developed principally as a research study. In 2003, CAL introduced a new computerized test named Basic Skills English Test (BEST) Plus. The aim is to test the ability in interpersonal language used in everyday life and cover communicative language functions ranging from providing personal information to

giving and supporting opinions. It has two versions as a computer-adaptive version and a semi- adaptive print-based version. In the computer version, the tester asks the test taker a question, listens to his/her answer, uses a rubric to give a score to the response, and enters the score into the computer. The computer program then selects the next test item, choosing questions most appropriate for the test taker's demonstrated ability level. Administering the test takes 5-20 minutes, which is a significant reduction in time compared to its predecessor, COPI.

Digital Video Oral Communications Instrument (d-VOCI) is a performance-based oral proficiency placement test developed by the Language Acquisition Resource Center at San Diego State University (LARC, 2003). The purpose is to use the advantage of multimedia-enhanced computer technology, that is the digital video delivery, to improve the previous face-to-face OPI by providing test takers with more control over various aspects of the testing situation and increasing the rater's efficiency in scoring the responses. Recently, the LARC at San Diego State University published LARCStar, a software that helps instructors to develop online language tests. Using the Internet as a test delivery medium, test takers complete the test by simply downloading the test items from the server to a computer anywhere the Internet is available and then, get immediate feedback.

The English Speaking Proficiency Test (ESPT) in Korea is developed to test English oral proficiency of test takers whose native language is not English. The test includes 25-26 questions and takes approximately 20 minutes to administer. At the beginning of the test, the test taker makes a self-introduction and then, several questions are asked such as describing pictures (e.g., weather, a map or a person) and acting to situations (e.g., a restaurant, a bank or a hospital) within an allotted time. Test takers are also required to read passages in which the aim is to assess their pronunciation and fluency. All test items are provided on a computer monitor and answers are recorded with a digital camera fixed on top of the monitor and a microphone placed in front of test takers. Responses are scored by trained human evaluators.

Purdue's Oral English Proficiency Test (OEPT) is developed for the purpose of screening non-native international English speaking teaching assistants' oral proficiency at Purdue University. The OEPT consists of 10 categories of test items including "short answer, personal history, read aloud, interpret graph, express an opinion, compare/contrast, offer advice, give information, and summarize casual speech"

(Purdue University, 2005). The test items are contextualized in academic environments (e.g. campus life, lectures, and so on). It takes approximately 45 minutes to administer the test. The responses are given to digital video prompts and recorded on a computer.

In SAM interview developed by Seyyed Abbas Mousavi (2007), it is aimed to study the impact of the use of digital video on test takers' speaking performance. It is a proficiency test measuring adult ESL learners' ability in spoken English regardless of any training they may have had. The test is delivered through the use of a multimedia package installed on a computer. The test items are provided in the form of video clips accompanied with relevant audio questions. Since the test is developed for university students, the content of the test items is designed as related to the test takers' future university life or social events (e.g. travel, environment, sport, general science, leisure activities, and libraries). SAM Interview is composed of five tasks. Three of them are at the level of the test taker's language proficiency determined by a self-assessment instrument and two of them were one level higher. It takes 15 minutes to complete these tasks. In assigning ratings, the International Second Language Proficiency Ratings (ISLPR) level description is used.

As can be seen from the studies above, computerized spoken language tests have been established in the past few years. When compared to face-to-face tests, language testers have reported both advantages and concerns about computerized speaking tests. The most prominent advantage of the use of computer-based speaking test seems to be their convenience and standardization of delivery, which promotes their reliability and practicality. In other words, their main advantages are "maximizing the efficiency of test administration" and "improving psychometric qualities of test scores" (Bachman, 1990 p. 336). As another advantage, once responses are digitalized on the server, a rater can review responses at any place and any time. The main concern about the use of computer-based speaking tests, however, is related to the inevitable narrowing of the test construct, since computerized speaking tests are constrained by the available computer technology and composed of limited test items which lack an interactional component. As a consequence, communicative language ability is not reflected in its breadth and depth and this deficiency creates potential problems for the construct validity of the test (Galaczi, 2010). In contrast, face-to-face speaking testing introduces the involvement of human interviewers and raters, which therefore brings about a broader test construct since interaction is an integral part of the test. The broader

construct, in turn, promotes the validity and authenticity of the test (Galaczi, 2010). The tradeoff, on the other hand, relates to the interlocutor variability; the potential influence of construct-irrelevant variance in the testing process. Each format in testing speaking ability bring along its strengths and weaknesses. Language testers can choose from fully automated speaking tests to ones involving human interviewers and raters, but the important point is 'fitness for purpose'. As emphasized in Galaczi (2010):

> "It is important to view computer-based and face-to-face speaking assessment as complementary and not competing perspectives, where technology is seen not as a replacement for traditional methods, but as a new additional possibility. Instead of viewing these different modes as one or another, they could be viewed as one and the other, and used in a way that optimizes their strengths and minimizes their respective limitations" (p.47).

Over the few years, computer based language testing has become prevailing worldwide. The number of institutions using computers as the main means of delivery for speaking tests has increased dramatically. Therefore, the increasing application of computer technology to the delivery of speaking tests has resulted in concerns about the appropriateness of the use of computers as a replacement for face-to-face speaking tests. In recent decades, studies have investigated the effects of test delivery mode on students' performances and compared direct and semi-direct methods of testing of second language speaking. The following part reviewed these studies and their findings.

### 2.6. Previous Research Findings

In comparing the effects of test delivery mode on the students' speaking performances, a number of research approaches have been adopted. These include assessing concurrent validity based on test scores (e.g. Stansfield, 1991; Stansfield & Kenyon, 1992; Kenyon & Malabonga, 2001), comparing the underlying constructs of test taker language output in different modes of tests (e.g. Zhou, 2015a, O'Loughlin, 2001; Shohamy, 1994) or collecting data on test taker feedback to assess face validity of the testing modes (i.e. Qian, 2009). In the present study, the concern was on two issues. The first issue was to compare the score equivalence of student performances between two test modes, the COPTEFL and the face-to-face speaking test. The second issue was to explore test taker attitudes to determine the appeal, or face validity of the COPTEFL. Research assessing concurrent validity and face validity, therefore, would be relevant to the scope of the current study and so, related research was investigated.

Research on the effects of delivery mode on testing speaking was mostly conducted on either tape-based test scores or on computer-based test scores. Despite

growing interest over the effects of computer delivery mode on test takers' speaking scores, a few research have been done on this issue (Zhou, 2015a). The most likely reason to this scarcity of research is that in recent years has the delivery of tests of speaking made possible through the use of computers. This section reviewed empirical research on the effects of test delivery mode, direct and semi-direct testing, for speaking assessment. It was divided into two parts as research on tape-based speaking tests and research on computer-based speaking tests.

### 2.6.1. Research on tape-based speaking tests

Previous studies on the comparability of speaking tests scores between tape-based speaking tests and face-to-face speaking tests have mostly focused on the OPI and the SOPI (Zhou, 2015) and relied highly on concurrent validation to confirm their equivalence (O' Loughlin, 1997). For example, Stansfield (1991) showed that scores on the OPI and SOPI in Chinese, Portuguese, Hebrew, and Indonesian had high correlations in the range of .89 to .95 and suggested based on the results that both measures tested the same abilities and therefore the SOPI could be reliably and validly used as a substitute for the OPI. For Stansfield (1991), the reason behind the high correlation may be because "neither format produces a 'natural' or 'real-life' conversation" (p.204). Even in a direct testing such OPI, he argues that both interviewer and test taker know that "it is the examinee's responsibility to perform. Little true interaction takes place" (p.205). However, in relation to such a conclusion, as O' Loughlin (1997) states further investigation needed: "[It] needs to be empirically investigated by examining data other than test scores (such as the discourse produced under the two test conditions) to discover, whether, in face, this is indeed the case" (p.63).

In a later study, Stansfield and Kenyon (1992) reviewed the studies that shed light on the comparability of the OPI and SOPI. These were the studies that conducted as part of the development of SOPIs in five less commonly taught languages (Chinese, Portuguese, Hebrew (USA), Hebrew (Israel) and Indonesian at the Center for Applied Linguistics (CAL). In their study, Stansfield and Kenyon (1992) discussed the results of these studies in terms of comparability of test scores and the agreement of scores obtained on the two tests. The same pair of raters scored OPIs and two or more forms of a SOPI taken by the same set of test takers. The results for the Chinese study ($n$= 32)

showed that the correlation between the OPI and SOPI averaged .98 across the two raters. In the Portuguese study ($n$= 30), the average correlation was .93. In the Hebrew (USA) study ($n$= 20), the average correlation was .94 and in the Hebrew (Israel) study ($n$= 20), it was .90. Finally as for the Indonesian study ($n$= 16), the average correlation was .95. Therefore, these findings revealed that when the same rater scored both the OPI and the SOPI, correlations were very high, averaging between .90 and .95. When different raters scored each test, that is, rater 1 scored the OPI and rater 2 scored the SOPI, or vice versa, the average correlations were also quite high. In the Chinese study, the average correlation between the two pairs was .90. For Portuguese, it was .93. For Hebrew (USA), it was .94 and for Hebrew (Israel), it was .90 and finally for Indonesian, it was .94. Thus, the findings again revealed high correlations between each pair of raters, averaging between .90 and .94, between the OPI score given by one rater and the SOPI score given by the other. In the study, Stansfield and Kenyon (1992) also examined the percentage of complete agreement between all OPI scores and all SOPI scores for all SOPI forms. Across the four Chinese SOPI forms, complete agreement between all pairs of OPI and SOPI scores awarded was 29%. It was 38% across three Portuguese SOPI forms. It was 60% and 43% across two Hebrew forms, Hebrew (USA) and Hebrew (Israel), respectively. Finally, it was 55% across two Indonesian forms. On the basis of the research that had been conducted on the SOPI and the OPI up to that time, Stansfield and Kenyon (1992) conclude that both test modes are highly comparable as measures of the same construct, oral language proficiency: “It is reasonable to claim that the OPI and the SOP1 are close enough in the way they measure general speaking proficiency that they may be viewed as parallel tests delivered in two different formats” (p.359).

As O’Loughlin (1997) states Shohamy (1994) takes a more skeptical position than the studies by Stansfield (1991) and Stansfield and Kenyon (1992) with regard to concurrent validation obtained through correlations between the scores on the OPI and the SOPI. She questions whether or not these high correlations demonstrate sufficient evidence to show that the two tests measure the same trait. By addressing her study in (1982), Shohamy (1994) states that

> “high correlations between tests provide necessary, but not sufficient, evidence for the appropriateness of test substitution. More convincing evidence for validation should be obtained by examining the specific language samples that two tests elicit and determining whether the tests are, in fact, the same” (p.102).

In comparing the mean score differences between the test scores of the OPI and SOPI, Shohamy (1994) shows that there was no significant difference in the mean scores of the 10 test takers. However, when compared the elicitation tasks for both tests, she found important differences in the range of language functions and topics. In the OPI, the oral language was a conversational interview and test-takers employed a variety of speech acts (i.e. elaborating and expanding responses, adding or requesting information, clarifying and, at times, interviewing the tester). As for topics, they were aligned in accord with the feedback received from the interviewer; thus, it was a dynamic process and there were signs of contextualization. As for the SOPI, the language was concise, with little elaboration beyond the specific task. After the task was given, test-takers provided two or three sentence answers in direct response to the task requirements and without adding any unelicited information. The language was of high lexical density, and it was more formal and cohesive than on the OPI. As for the topics, there was a sharp shift from one topic to another. After completion of a task, test takers were ready to continue. There was frequent paraphrasing and parts of questions were sometimes repeated in the answers. The general implication in Shohamy's (1994) research is that even though there exists a high correlation between scores on the two tests, the OPI and the SOPI do not demonstrate to measure the same language construct and therefore, they are not necessarily interchangeable as tests of oral proficiency.

Kenyon and Tschirner (2000) compared test scores from the German OPI and SOPI taken by a 20 test takers and the findings revealed no statistical mean difference between the two tests and high correlations ($r$= .96) as well as high percentage of absolute agreement (90%) between ratings of the two test. They concluded that as with inter-rater agreement, there exist high correlations between the OPI and SOPI. However, this does not necessarily imply perfect agreements between performances on each test.

In this section, the studies comparing test scores between tape-based speaking tests and face-to-face tests were reviewed. Their findings have supported the score equivalence consistently. However, as (O'Loughlin, 1997) states high correlations only imply that we can predict one test score from another. Correlation is a measure of the association of two tests (Fulcher, 2003). It is still possible that test takers may have performed better on either test. In the following section, the studies comparing test

scores between computer-based speaking tests and face-to-face speaking tests were reviewed in terms of (1) score comparability and (2) face validity.

### 2.6.2. Research on computer-based speaking tests

Recent advances in computer technology have promoted computer delivery of speaking tests. Applying computer technology to delivering tests now appears to be a popular practice (Qian, 2009) and recent research into this issue has begun to empirically evaluate the effects of computer delivery mode on speaking tests.

Kenyon and Malabonga (2001) explored the attitudinal reactions of test takers to taking different formats of oral proficiency assessments in Spanish, Arabic, and Chinese. A total of 55 students participated in the study. They were taking courses at universities in the Washington, DC metropolitan area. The participants ($n$=24) taking the Spanish tests were administered three types of tests: a tape-mediated Simulated Oral Proficiency Interview (SOPI), a Computerized Oral Proficiency Instrument (COPI), and the face-to-face Oral Proficiency Interview (OPI). The participants taking the Arabic ($n$=15) or Chinese ($n$=16) tests took the SOPI and the COPI only. The correlation between the test takers' scores in the SOPI and the COPI was .95, in the COPI and the OPI, it was .92, and the SOPI and the OPI scores correlated at a level of .94, which indicated that the test takers scored very similarly across the tests. After the administration of each test, test takers completed a questionnaire on their attitudes towards and perceptions of that test, and finally they were given another questionnaire asking them to compare the test formats. Comparisons were made in six categories: opportunity to demonstrate strengths and weaknesses in speaking, test difficulty, test fairness, nervousness, clarity of instructions, and representativeness of the performance. The results showed that both the SOPI and the COPI were perceived as equally fair and clear. However, SOPI was found more difficult. When the perceptions in relation to the semi-direct format (COPI/SOPI) were compared to those related to the OPI, it was revealed that there was still a definite preference for the OPI as the test takers reported that it provided a better opportunity to demonstrate their speaking ability. Some of the test takers reported that many aspects of their speech communication could be captured in a technology-enhanced assessment (i.e. pronunciation, vocabulary, fluency, grammatical control, and even the ability "to think on the spot"). However, they strongly felt that the technologically-mediated tests could replicate the interactional

nature of a conversation of the face-to-face test.

To study the effect of the use of digital video on test takers' performance, Mousavi (2007) developed a computer-based test (CBT) (SAM Interview). SAM Interview was a proficiency test in the sense that it was developed to assess the test-takers' ability in spoken English regardless of any training they may have had. Test tasks were designed on the basis of the video content at three difficulty levels (pre-intermediate, intermediate, and advanced). SAM Interview was administered to a group of ESL (English as a Second Language) learners ($n$= 30) and their performances were recorded for grading. The test-takers' performance on the SAM Interview test and on IELTS (International English Language Testing System) were correlated in order to find out whether CBT and face-to-face speaking test scores comparable. The Spearman *rho* correlation coefficient between the test-takers' performance on the SAM Interview test and on IELTS or ISLPR was found to be .63, which was considered a high moderate index of reliability. For the research, a post-test feedback questionnaire was given to the test takers to explore their reactions and attitudes to newly developed SAM Interview. Analysis of the results demonstrated that the test takers had a highly positive attitude towards digitally delivered test prompts compared to the face-to-face oral proficiency tests. Also, for test takers, computer-based tests were less threatening and more comfortable than face-to face oral proficiency tests.

In an attempt to explore the interactions between two test taker characteristics (i.e. self-reported English study period and self-reported computer skills) and test delivery format during spoken English proficiency assessments, Yu (2006) administered a computerized spoken English test, an audio-taped SPEAK (the Speaking Proficiency English Assessment Kit) test, and a questionnaire to a total of 210 international students whose native language was not English. The SPEAK test was developed by the ETS. This test assesses oral proficiency of non-native English speakers world-wide since it was first published in the early 1980s. Two trained raters awarded ratings on the participant response samples. Out of a total of 210 response samples, 25 samples were randomly selected. The two raters individually evaluated the 25 comparison response samples. A high correlation coefficient of .99 was found between the test scores for the 25 response samples. The main data sources included the results of test scores from two test delivery formats and replies to a questionnaire. Test delivery format was represented with two formats: a computerized spoken English test and an audio-taped

spoken English test. The number of self-reported years of English language study was self-reported on two levels: less self-reported English study and greater self-reported English study. Computer use was self-reported on two levels: less computer use and more computer use. For data analysis, a three-way ANOVA was employed. The results showed that the interaction among all three independent variables (i.e., self-reported years of English study, self-reported computer use, and test delivery format) was not significant. However, self-reported years of English language study and test delivery format cooperatively produced a significant influence on test scores for the spoken English test. Specifically, it was revealed that the computerized speaking test, not the audio-taped SPEAK test, appeared to have impact on test results more for the group that self-reported less English study than for the group that self-reported greater English study. Also, self-reported computer use did not significantly affect test results during oral proficiency assessments.

Jeong (2003) investigated the relationship between Korean college students' electronic literacy, measured through the multimedia-based English oral proficiency interview utilizing d-VOCI (digital-Video Oral Communication Instrument) and their multimedia-based OPI scores, where the test takers were asked to give responses to the prompts given by a computer and record their voices. The subjects were administered both a face-to-face and a multimedia-based English oral proficiency interview. The data from SEM analysis showed that there was a moderate positive relationship between test takers' electronic literacy and their English oral proficiency (.44). However, the reliability of the two English oral proficiency tests used in this study (face-to-face and computer-based format) revealed only a weak relationship (.30). Approximately, 70 % of the subjects indicated that they preferred the face-to-face interview, and 30 % reported that they preferred the d- VOCI. Among those whose preference was for the face-to-face interview, more than 80 % stated that the face-to-face interview seemed to be the better test method since it was "real" and "interactive." The students who indicated a preference for d-VOCI were not found to get particularly high or low scores from d- VOCI or the electronic literacy survey. In other words, neither English proficiency nor electronic literacy appeared to have an impact on students' preference between the d-VOCI and the face-to-face interview.

In an investigation of the attitudinal reactions of test takers to two test delivery modes, a computer-based and a face-to-face mode, Joo (2008) explored Korean

university students' attitudes to a Face-to-Face Interview (FTFI) and a Computer Administered Oral Test (CAOT) first and then their performance on the tests, and finally their effects on performance on the two tests. Both tests were administered to a total of 42 university students. After taking the tests, the participants completed a questionnaire about their attitudes towards both tests and their perceptions of the tests. Among them, ten students were interviewed after the questionnaire to better understand their attitudes and performance. As for the results, the inter-rater reliability findings showed reasonably strong Spearman correlations of r=0.77 for the FTFI and r=0.87 for the CAOT. In terms of performance, a mismatch was found between the difficulty of the tests and the ability of the test takers. It was revealed that the tests were very challenging and hard for the test takers. The raw scores ranged only from 1 to 5 on a rating scale of 7 (mostly used 1 to 3) and this results suggested the test takers' low speaking abilities. For the investigation into the equivalence of two test modes, the Spearman correlation between the logit scores obtained from the two tests was computed and it was found as $r$ =.71, which was stated as not very high compared with previous studies on direct and semi-direct tests. For further understanding, a chi-square test was run to compare the test taker ability estimates obtained from the scores on the FTFI and the CAOT. The results indicated that the two tests were not the same in terms of their difficulty: chi-square 17.1 (p=.00) with separation 2.75 and reliability .88. In terms of attitudes towards test formats, the results of the study showed that Korean university students had much more favorable attitudes to the CAOT compared with previous studies on direct and semi-direct tests, but they still preferred the FTFI to the CAOT in spite of significant negative attitudes to the FTFI in relation to aspects such as nervousness, preparation time, and tiredness.

In an attempt to compare two testing modes, face-to-face and computer formats by analyzing reactions and perceptions of a group of test takers, Qian (2009) conducted a study in the context of a university setting in Hong Kong ($n$= 186). Data were collected through two oral English proficiency tests: (1) face-to-face format was represented by the speaking component of the International English Language Testing System (IELTS), and (2) computer format was represented by the speaking component of the Graduating Students' Language Proficiency Assessment–English (GSLPA). The findings of the study showed that the majority (57.6%) of the test takers actually did not show a particular preference with regard to the testing format. Even though a large

proportion of the test takers showed no particular preference with regard to the testing format, the number of participants who strongly favored face-to-face format far exceeded the number strongly favoring computer format. The participants' main reason cited for not favoring GSLPA was its inability for the interviewer and test taker to interact during the test, which seemed to have created a psychological barrier for the test takers.

In their study, Kiddle and Kormos (2011) investigated the effect of mode of response on test performance and test-taker perception of test features by comparing a semi-direct online version and a direct face-to-face version of a speaking test. They conducted the research with 42 participants from a Chilean university. For the purposes of the study, two versions of the speaking test were developed in parallel to enable maintain equivalence with regard to test content. The online test was designed using the Questionmark Perception v4.4 platform, which was reported as a sophisticated test-delivery platform. In the face-to-face version of the test, test takers were greeted by the interviewer and sat facing him or her across a desk, with a laptop computer facing the test taker on the desk. The interviewer began the test by displaying the same Task 1 instructions as in the online test, in English and Spanish, on a piece of paper in front of the test taker and repeating the same instructions as in the online Task 1 video. When the test takers had finished speaking, the interviewer repeated the procedure for the remaining tasks. To provide further insights into test takers' perceptions of the two test delivery modes, the researchers also asked test takers to complete a questionnaire after sitting each version. The findings revealed that the Cronbach alpha coefficient for the marking of computer-based test scores was .84 and for the face-to-face version was .85, which were reported as within the acceptable range. A series of independent sample *t* tests were run to find out whether the test order had any effect on the results. The analyses showed that the order was not deemed to be a significant factor. To determine the effect of the mode of response on test scores, a series of paired- sample *t* tests were employed to measure the significance of differences between means in the total test scores. The analysis indicated that there was a significant and strong correlation between the two sets of scores ($r = .85$), and the *t* test shows no significant difference between means on the two versions of the test. Considering test taker perceptions, it was revealed that test takers were generally found to be favorable with regard to both versions of the test. The mean scores on the relevant questionnaire items showed that

test takers agreed that both the online and the face-to-face versions of the test were fair and that preparation time was sufficient under both conditions. However, a significant difference was found in terms of perceptions of fairness. Accordingly, test takers did not fully accept the computer-based test as equivalent to the face-to-face test.

Öztekin (2011) aimed at reflecting on the use of computer-based oral assessment as a substitute for a face-to-face speaking test at a Turkish tertiary school at pre-intermediate and intermediate levels. It explored the relationships between test scores obtained in two different test modes at two different proficiency levels, the test takers' perceptions of the test modes, and their anxiety levels in relation to speaking in a foreign language, speaking tests, and using computers. The study was conducted at Uludağ University School of Foreign Languages and a total of 66 students and four instructors of English participated in the study. Data were gathered through four computer assisted and four face-to-face speaking tests, a questionnaire on Computer Asssisted Speaking Assesment (CASA) perceptions and another on Face-to-face Speaking Assessment (FTFsa) perceptions, a speaking test and speaking anxiety questionnaire, and a computer familiarity questionnaire. In relation to the perceptions of the test modes, the findings showed that the test takers were found to prefer the FTFsa over the CASA and had more positive feelings towards the former in general. When the test-mode-related anxiety subscales of the two perceptions questionnaires were explored, the test takers were found to be more anxious in the CASA than in the FTFsa at both levels. The analyses employed to examine the relationship between speaking anxiety and speaking test anxiety, and the test-mode specific perceptions showed that there was a significant negative correlation between speaking anxiety and the CASA perceptions at the pre-intermediate level. No significant correlation was revealed between speaking anxiety and the FTFsa perceptions or the test scores the pre-intermediate test takers obtained in either test mode. At intermediate level, a significant negative correlation was revealed between speaking anxiety and CASA and FTFsa perceptions. Finally, as for the findings related to the computer attitudes and their relationship with the perceptions of the test modes and test scores the results revealed a marginally significant correlation only between pre-intermediate level test takers' computer attitudes and their CASA perceptions (a negative correlation) in addition to their CASA scores (a positive correlation).

In a recent study, Zhou (2015a) investigated the effects of delivery mode on the

testing of second language speaking. In the study, the test scores and the underlying factor structures of monologic tasks were compared between computer delivery and face-to-face modes. Seventy-nine Japanese students participated in the study. Two raters scored on an analytic scale of 1 to 4 for each of the four rating elements: grammar, vocabulary, fluency, and pronunciation. Participants were awarded ratings from a different pair of raters. In order to measure the degree of consistency between the ratings, two types of rater consistency indexes were calculated: Pearson' s correlation and rating agreement. The Pearson' s correlation coefficients were from .52 to .75 for the computer-delivered tasks and .60 to .74 for the face-to-face tasks. The rating agreements for both delivery modes were satisfactory with moderately high exact agreement (54.4%– 68.4%) and adjacent agreement (29.1%– 45.6%) for the face-to-face tasks; similarly moderate exact agreement (49.4%– 72.2%) and adjacent agreement (26.6%– 40.5%) were found for the computer delivered tasks. One-way multivariate analysis of variance results revealed no significant differences between the test scores from two delivery modes. Exploratory factor analysis also did not reveal differences in the underlying factor structures of the two delivery modes. In relation to the results, Zhou (2015a) stated that the findings provided evidence for the use of monologic tasks in computer-delivered speaking tests.

Thompson, Cox and Knapp (2016) investigated the effect of test method on oral proficiency scores and student preference by comparing two test formats: Oral Proficiency Interview (OPI) and Oral Proficiency Interview–Computer (OPIc). 154 Spanish language learners at a large private university in the western United States participated in the study. Participants were administered both the OPI and OPIc within a 2-week period using a counterbalanced design. In addition, they took both a pre- and post-survey that obtained data about their language learning background, familiarity with the OPI and OPIc, preparation and test-taking strategies, and evaluations of each test. The results revealed that both groups got higher scores on the OPIc than they did on the OPI, demonstrating that there was no ordering effect. Furthermore, 54.5% of the test takers were awarded the same rating on the OPI and OPIc, with 13.6% of test takers scoring higher on the OPI and 31.8% scoring higher on the OPIc. While the findings revealed that test takers scored significantly better on the OPIc, the overall effect size was quite small. The study also found that the overwhelming majority of the test takers preferred the OPI to the OPIc. In the majority of cases, test takers reported that OPI was

more natural, more enjoyable, and more like a real speaking test. They also stated that the OPI was a better measure of their proficiency. However, those who preferred the OPIc indicated that talking to a real person increased their level of anxiety and made them feel that they were being judged for their mistakes. They reported that the computer's nonjudgmental presence and the opportunity to have tasks repeated multiple times helped them to be more relaxed and therefore, better able to show their actual speaking ability.

The literature on the comparability of direct and semi-direct speaking tests have focused largely on correlations or analyses of test outcomes/products including test scores, underlying constructs of test taker language output in different modes of tests, and test taker reactions (i.e. attitudes). 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 (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.

This section reviewed the empirical research into the comparability of direct and semi-direct tests of oral proficiency. When we look at the current literature on the semi-direct language testing, we see that over the few years, computer based language testing has become prevailing worldwide. Latest developments of language testing have led to incorporating computers in testing language proficiency due to factors such as economy and task versatility, and consequently, the number of institutions preferring computers as the main means of delivery has increased dramatically. Overall, computers appear to fit properly with the new demands of modern and academic oriented higher education test constructs (Garcia-Laborda, Santiago, de Juan & Alvarez, 2014). Such as, the PAULEX project, a web-based assessment system for foreign language exams in Spain, indicated a high degree of acceptance by its users (de Siqueira, Martínez-Sáez, Sevilla-Pavón, and Gimeno-Sanz, 2011). According to Garcia-Laborda, Magal Royo and Barcena Madera (2015), countries like the United States or the United Kingdom are spearheads of these practices. However, as stated in their study, this situation is not the same everywhere. Current research in computer-based language testing in Turkey is rather limited to just a few research. The following section reviewed the language testing studies conducted in Turkey.

## 2.7. Language Testing Studies in Turkey

In order to find out the studies on language testing field in Turkey, Turkish national research archives -where the researchers make an entry for their studies are used. These archives are ULAKBIM (http://arsiv.ulakbim.gov.tr/), YÖKAKADEMİK (http://akademik.yok.gov.tr/AkademikArama/) and Ulusal Tez Merkezi (https://tez.yok.gov.tr/UlusalTezMerkezi/). Only the studies published after the year of 2000 are included. According to the observations, language-testing studies in Turkey have covered the following issues (see Table 2.6):

**Table 2.6.** *Language testing studies in Turkey (2000-2017)*

| Issue | Year | Author(s) | Title |
|---|---|---|---|
| **Perceptions** | 2016 | Ağçam & Babanoğlu | Students' Perceptions of Language Testing and Assessment in Higher Education |
| | 2013 | Gönen | Teachers' Perceptions of Classroom Based Language Assessment in Tertiary Level English Language Programs in Turkey |
| | 2012 | Şahinkarakaş | The Role of Teaching Experience on Teachers' Perceptions of Language Assessment |

| | | | |
|---|---|---|---|
| | 2007 | Ceylan | Ethical Concerns in Language Testing at University Level |
| | 2017 | Aydın, Geridönmez, Polat & Akay | Feasibility of Computer Assisted English Proficiency Tests in Turkey: A Field Study |
| | 2016 | Aydın, Akay, Polat & Geridönmez | Proficiency Exam Procedures in Turkish Universities and Attitudes Towards Computerized Language Assessment |
| **Further insights** | 2014 | Cephe & Toprak | The Common European Framework of Reference for Languages: Insights for language testing |
| | 2011 | Öztekin | A Comparison of Computer Assisted and Face-to-Face Speaking Assessment: Performance, Perceptions, Anxiety, and Computer Attitudes |
| **Factors** | 2012 | Engin | Variation of Scores in Language Achievement Tests according to Gender, Item Format and Skill Areas |

a. Perceptions:

The studies exploring perceptions on language testing have been carried out with a focus on perspectives of students or teachers with regard to language testing. Ağçam and Babanoğlu (2016) was one of the recent studies that was motivated to explore the perceptions of students on foreign language (FL) assessment in higher education in Turkey. It primarily investigated the students' perceptions of core language skills, assessment types and question types used in assessing their FL development/ proficiency during an academic year. The findings of the study showed that most participants found assessment necessary in their foreign language education, and that speaking and listening were regarded as the most important skills, while grammar and reading were considered the least important. In Şahinkarakaş (2012), the study focus was on teachers and the aim was to explore how language teachers conceive language assessment and whether these conceptions differ according to teaching experience. In order to reveal participants' conceptualization for language assessment, prospective language teachers and language teachers with different years of teaching experience were asked to describe 'language assessment' using a metaphor. Analysis of the metaphors obtained by the written descriptions of the participants revealed two main conclusions. Teachers, in general, considered assessment as being embedded within instruction. In other words, they regarded assessment as a way to offer evidence of teaching and learning. The other conclusion was related to teacher self-efficacy that prospective teachers seemed to have the highest self-efficacy level. In relation to the experienced teachers' self-efficacy, Şahinkarakaş (2012) had two possible suggestions: (1) experienced teachers valued themselves too highly and therefore did not need to

provide evidence of their teaching effectiveness, or (2) experienced teachers lost their motivation for effective teaching. In the studies by Gönen (2013) and Ceylan (2007), the attempt was to investigate the perceptions of teachers in relation to language testing. Gönen (2013) explored the perceptions of English Language Teaching (ELT) Department instructors on ethics concept in language testing. The study further aimed at collecting information in order to develop a 'code of ethics' or 'code of practice' in language 'testing for Turkish Education System. The findings of the study revealed that usually the teachers hold and exercise their own firm beliefs in terms of classroom based language assessment. The teachers reported that in some cases they did not have opportunity to put their beliefs into application, they had intensive syllabus to finish in a very limited time. Ceylan (2007) investigated teachers' perceptions and their application with relating to classroom based language assessment. It primarily examined the teachers' current operational principles of assessment and their practices in higher education in Turkey. The research was conducted with the instructors of ELT Department at Çanakkale Onsekiz Mart University. The results showed that most of the instructors did not know whether there was a 'code of ethics' for language testing in Turkey or not. However, they were found to believe that a 'code of ethics' for language testing was necessary and indicated that 'a code of ethics' should be developed considering the value systems of the Turkish culture.

b. Further insights:

In order to provide insights into language testing practices, Aydın, Akay, Polat and Geridönmez conduct a study in which the aim was twofold: (1) to explore the practices offered by the universities to prepare reliable and valid language proficiency exams and to discuss the feasibility of these practices in their contexts; (2) to gather opinions from state universities in Turkey about the use of computer assisted assessment techniques in the evaluation of language proficiency, and to identify the existing practices if there are any (2016). One year later, they carried out another study in which the researchers collected information on technical and procedural aspects of the administration of international computer based tests by conducting interviews with two of the internationally recognized testing organizations abroad - namely, Educational Testing Service (ETS) and Pearson (Aydın, Geridönmez, Polat, Akay, 2017). By reporting the findings from the interviews conducted with the professionals in the testing organizations visited, the researchers aimed to suggest a possible path to develop

Turkish nationwide computer based proficiency test. Related to computer-based assessment, there was another study, which was a master thesis, conducted by Öztekin (2011). The study intended to reflect on the use of computer mediated oral assessment as a substitute for a face to-face format at a Turkish tertiary school at pre-intermediate and intermediate levels. It aimed to shed light on the degree to which computer mediated oral assessment was valid and equal to the face-to-face format, the test takers' perceptions of oral assessment, and the relationship between the anxiety levels of the test takers and the test mode. This study was further explained in section (2.6.2). In their study, Cephe and Toprak (2014) attempted to investigate the practical concerns and potential problems with regard to the CEFR in terms of language testing and to discuss some practical implications for language testers and language teachers in terms of test generation and alignment. They stated that the CEFR can be used to determine the objectives for teaching and assessment but inadequacies also prevail. By referring to Weir (2005), they explained the deficiencies of the CEFR that it "does not help to develop comparable tests let alone helping us to decide if these tests are comparable" (p.86).

c. Factors:

In her master thesis, Engin (2012) investigated the variation of scores in language achievement tests. This study examined how language learners scores in language achievement tests vary according to such factors; gender, item format (matching, fill in the blanks, find the correct form, multiple choice, open ended, and paragraph writing) and skill areas (reading, writing, listening, grammar, and vocabulary); and it also examined whether the male and females score vary according to item format and skill areas. The results of the study demonstrated that gender did not have a significant effect on the total scores of the students in language achievement tests. However, students' total scores were found to vary significantly depending on both the item format and skill areas in the test. Males' and females' mean scores also were found to have differences depending on both item format and skill areas. According to the analyses of the findings, females outperformed males significantly in two item formats; 'find the correct form' and 'paragraph writing' questions, while males did not show any superiority in any item format. Also, in skill areas, females outperformed males in three skill areas; 'writing,' 'grammar' and 'vocabulary' whereas males scored higher only in one skill area; 'listening.'

As stated at the beginning of the section, the number of the studies was restricted to the entries in the archives. The primary reason to this restriction was that there were huge amount of studies in the literature when searched with a 'language testing' key word and 'language testing in Turkey' key word reveals lots of irrelevant findings, which were not related to the studies in Turkey. Since aforementioned archives provided the titles of studies by almost each academician in Turkey, the researcher made uses of them.

When we look at research in language testing in Turkey, we see that it was rather limited to just a few research, but research in computer-based testing was almost non-existing. Since 2011, there has been an interest from the perspective of research about the use of computers in language testing but how computer-based language tests can be developed and validated for the purposes of uses has not been studied, rather its importance was emphasized (Aydın et al. 2016; 2017).

It is clear that transition to computer-based language testing is an important step for institutions. The results of the studies by Aydın et.al (2016, 2017) revealed the possibility of a transition from paper based test system to a computer based one in Turkey. In the findings of these studies, the solution to the problems experienced in testing the proficiency level of language learners at the preparatory programs was stressed by establishing a computer-based testing system from which all institutions could benefit. However, as pointed out in these studies, jumping into a system without considering necessary precautions would bring more problems rather than solving the existing ones. In order to start such a transition process, this study addressed the difficulty in testing oral proficiency at the preparatory programs in Turkey, and attempted to make necessary planning from pre-planning to assessment by considering all the stages of developing a computerized oral proficiency tests. What decisions were made to prepare computerized oral proficiency tests and how these decisions were planned to develop, administer and validate such a test were explained in the next section (see Methodology Chapter).

# 3. METHODOLOGY

## 3.1. Introduction

This chapter outlined the methods and procedures of the present study. The following subsections reviewed the research design, setting and participants, instruments, data collection procedure and data analysis.

## 3.2. Research Design

This study is about the development and validation of the Computerized Oral Proficiency Test of English as a Foreign Language (COPTEFL). After the development of the COPTEFL, the study attempted to investigate the extent to which students perform differently from the face-to-face speaking test. That is, whether students' performance differed when talking to a computer from when facing an interviewer. It is important to examine this issue since it relates to the fundamental questions of test validation, i.e. to ensure the score validation (Zhou, 2008). In order to investigate the equivalence of test scores across mode in a more comprehensive manner, the study also gathered information on test takers' attitudes towards both test conditions. With these aims, the study addressed the following research questions:

1. What are the cognitive demands of the speaking test tasks?
2. What are the contextual characteristics of the speaking test tasks?
3. 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?
4. What are the attitudes of test takers in relation to the test delivery modes?

In order to investigate the research questions listed above, the current research employed a "mixed methods research design" (Creswell, 2012). A mixed methods study involves the collecting, analyzing and "mixing" of both quantitative and qualitative data in a single study. The reason of combining methods was to achieve a better understanding of the comparability of the scores from two modes of testing than either method by itself. Within the types of mixed methods design, this study used "the explanatory sequential design" (Creswell, 2012, see Figure 3.1).

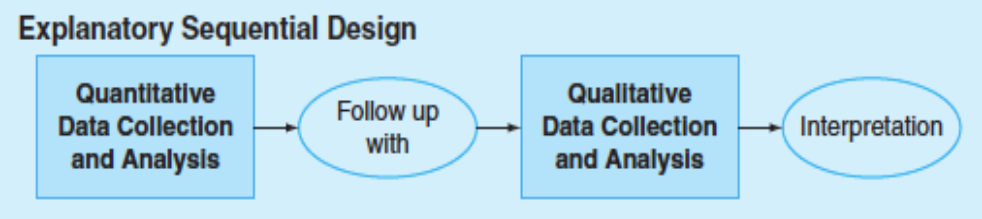



**Figure 3.1.** *Explanatory sequential design scheme (Creswell, 2012, p.541)*

This type of design consists of first collecting quantitative data and then collecting qualitative data to explain on the quantitative results (Creswell, 2012). The rationale employing the explanatory sequential design in the present study was that the quantitative data; test scores across modes, provided a general picture of the research problem; whether the scores obtained from the administration of the COPTEFL were comparable with the conventional face-to-face speaking test, and more analysis, specifically through qualitative data collection was needed to elaborate on the comparability of the scores. By looking at the research problem from a different angle; whether test takers' attitudes showed that the COPTEFL system seemed to them to test the same ability the face-to-face speaking test tested, it was aimed to expand the scope and breadth of the study.

### 3.3. Setting

The study was conducted at Anadolu University School of Foreign Languages (AUSFL), Turkey. The school is a preparatory program which aims to equip the students with necessary language skills in order to follow the academic education in their departments. It provides extensive English language education totalling 22-24 hours per week. There are approximately 3,000 students taking this intensive language instruction each year (see https://ydyo.anadolu.edu.tr). The students who are registered for the program have to take two proficiency tests. One of them is given at the beginning of the term in order to put the students into groups based on the scores they got and the other one is given at the end of the term as an exit exam measuring if they have reached a certain proficiency level. 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.

From its establishment year, 1998 to 2010, the school employed a skills-based language-teaching program of foreign language instruction (Aydin, Unver, Alan &

Saglam, 2017). In 2010, the school shifted from a skills-based language teaching program to an integrated skills program and the curriculum aligned with the CEFR-Common European Framework Reference proficiency levels. 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 2014-2015 academic years, the school moved away from CEFR towards The Global Scale of English (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 was designed based on the GSE 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 because "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 (see Appendix A).

### 3.4. Participants

This part specifies what sort of learner the test was designed for. This test was designed for newly arrived students in preparatory schools of university who have finished their high school or the students who are studying in preparatory schools of university and have to take exit exam to demonstrate that they have gained the sufficient proficiency in English for academic study in their departments. The participants of the study were 45 non-native speakers of English, who were enrolled in the preparatory program at the time of data collection. Participation was on a voluntary basis. The

students who had interest in assessing their speaking skill and willingness to try a computer-based spoken English test were contacted to ask whether they could participate in the present study by mail. Those who were enrolled were randomly assigned to take either the COPTEFL first and then the face-to-face speaking test, or vice-versa.

### 3.4.1. Test development team

#### 3.4.1.1. *Item writers and raters*

There were 8 item writers in the study. Item writers were also the raters. They are professional instructors who work full-time for 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. Their experience provided insights into what such students find easy and difficult, which item could work or not, and what kind of problems we could face with in rating.

#### 3.4.1.2. *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 rating. Another one was experienced in teaching speaking and language testing in our foreign languages department. The others were subject experts, the researcher was one of them.

#### 3.4.1.3. *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. The researcher helped developers in designing the system so that the COPTEFL system would work and look like as it was planned.

### 3.5. Instruments

The data for this study were collected from multiple sources, including (1) the COPTEFL scores, (2) the face-to-face speaking test scores, and (3) open-ended questions comparing test-taker attitudes towards the two testing modes.

#### 3.5.1. The COPTEFL

##### 3.5.1.1. *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 Web as its delivery medium. It was designed to offer users an alternative to face-to-face oral proficiency tests. In this project, the aim was to increase the efficiency of test delivery and administration and to provide a unified and standardized 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 permissions to access (see Figure 3.2).

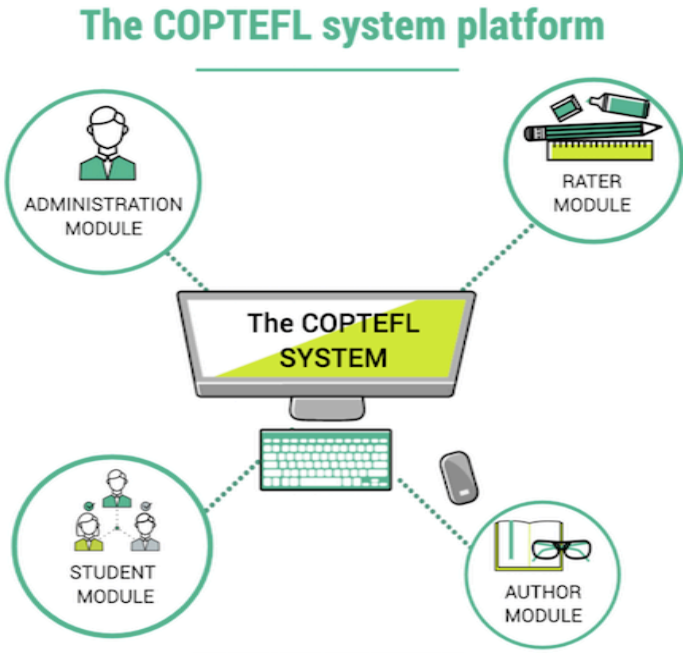



**Figure 3.2.** *The COPTEFL system platform*

In order to keep the system reliable, only the administrator(s) have access to other modules. Each module and their 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 to the data of a user in particular, editing users' personal data, setting of user deadlines (see Figure 3.3).

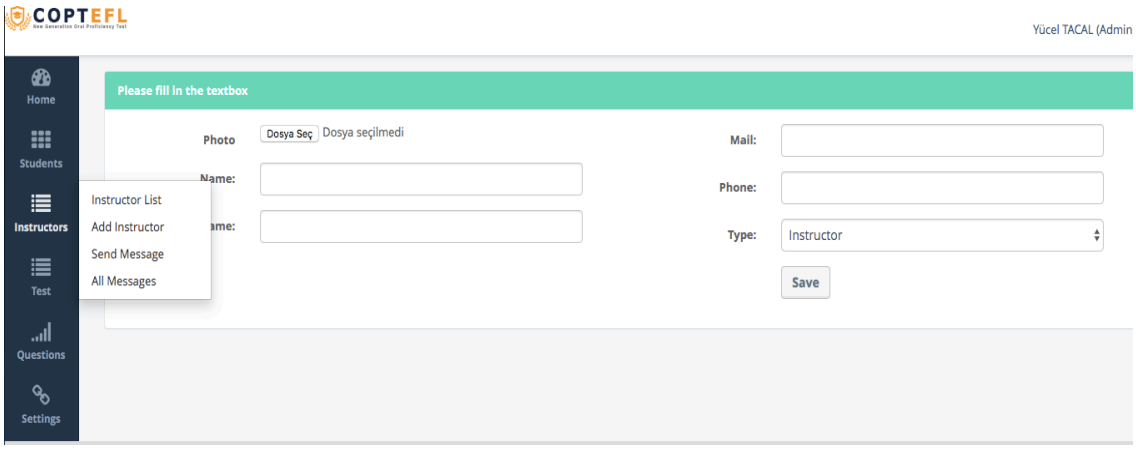



**Figure 3.3.** *Administration module: Managing users*

(b) Setting the time for the test and its announcement to the student module for test taker registration (see Figure 3.4).

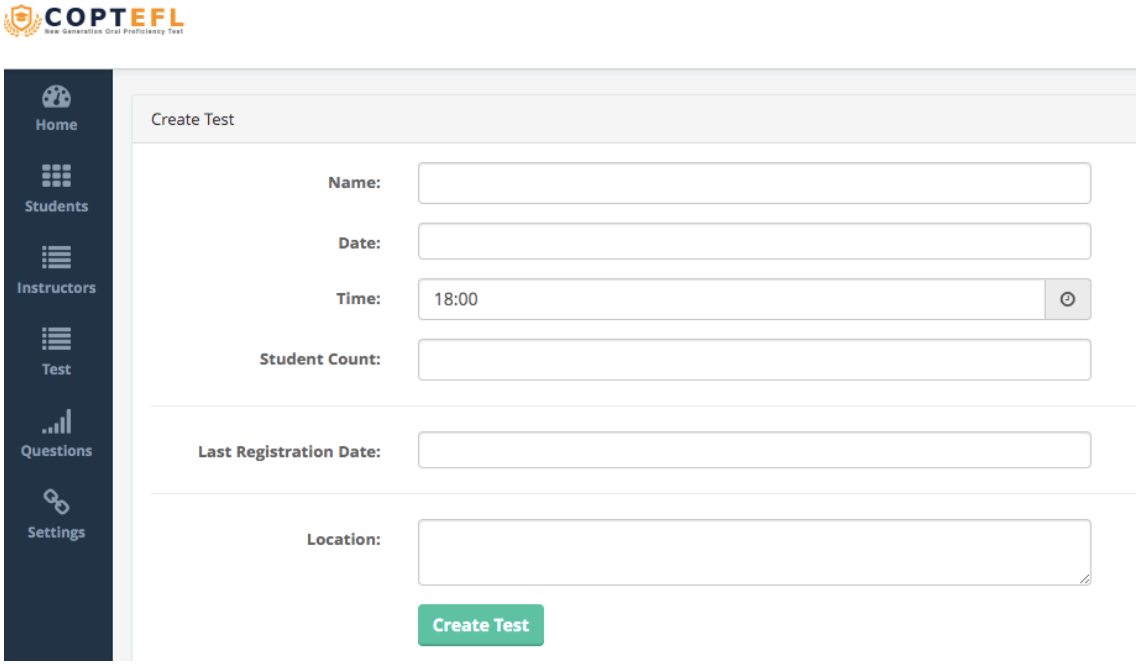



**Figure 3.4.** *Administration module: Setting time for the test*

(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 into the item pool by the system automatically (see Figure 3.5).

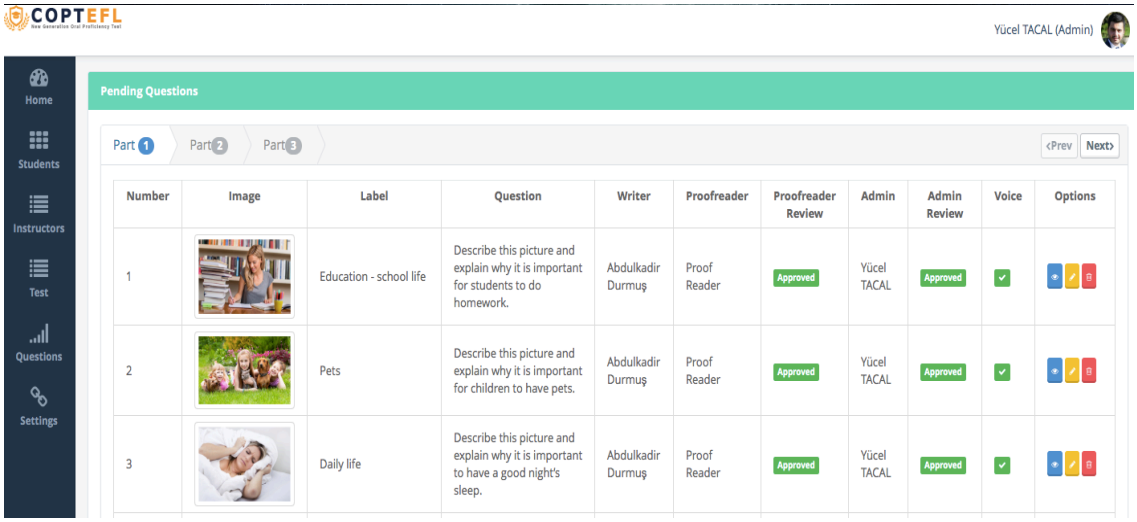



**Figure 3.5.** *Administration module: Control panel for the written test items*

(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 to the students' answers to the questions during the test administration

(i) Accessing to the test results (see Figure 3.6).

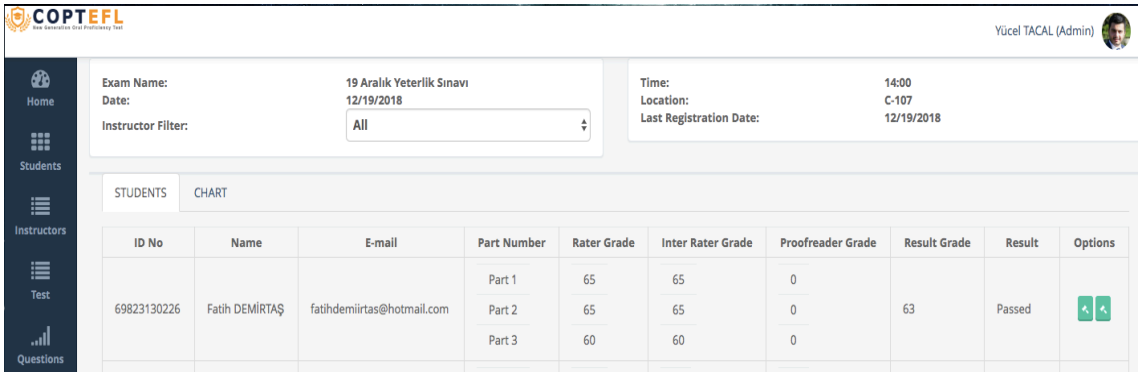



**Figure 3.6.** *Administration module: Accessing to the test results*

**2. Author Module:** This is the module used by the item writers to develop and edit tasks for test item pool (see Figure 3.7).

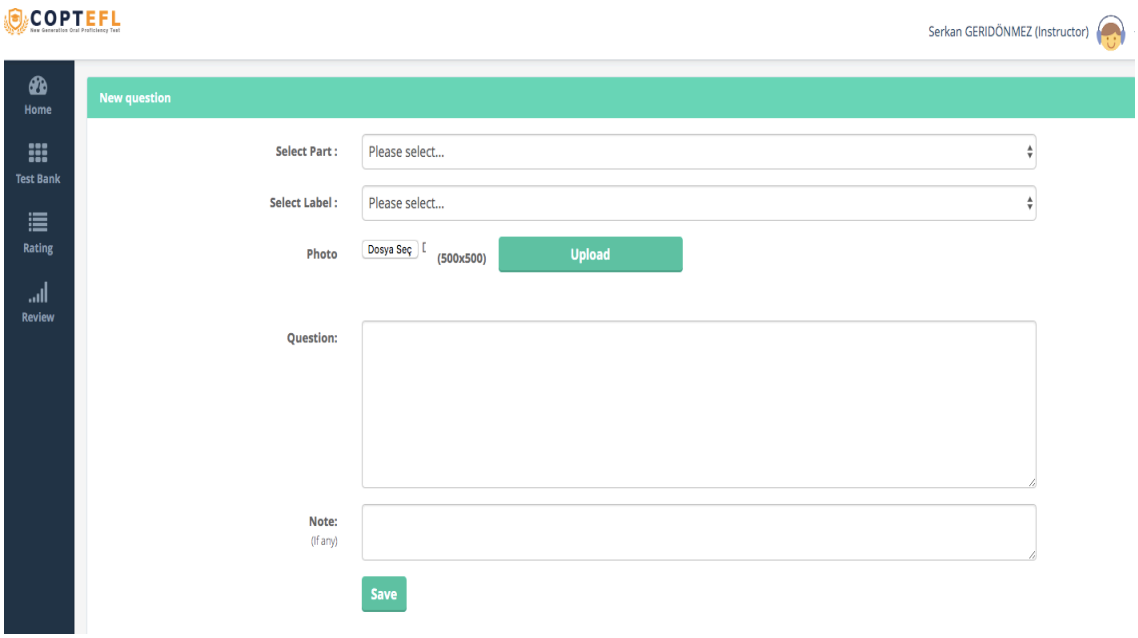



**Figure 3.7.** *Author module: Adding questions*

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 the 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 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. The tasks are scored according to the following assessment criteria: Pronunciation, Fluency, Grammar range and accuracy, Adequacy of vocabulary for purpose and Task fulfillment. Each criterion is shown with its explanation on the page and raters see criteria section while they are listening to the answers. They can give scores at the same time they are listening to (see Figure 3.8).

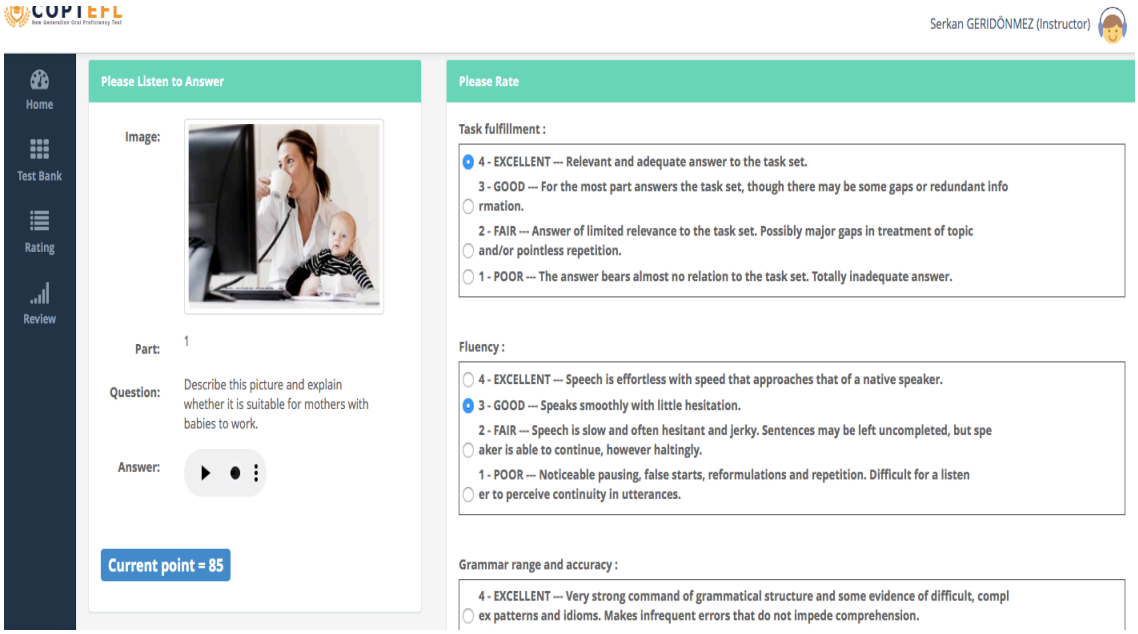



**Figure 3.8.** *Rater module: Giving scores*

When they complete marking, the computer shows the total grade that a student gets after scoring process and then the rater can submit the score (see Figure 3.9).

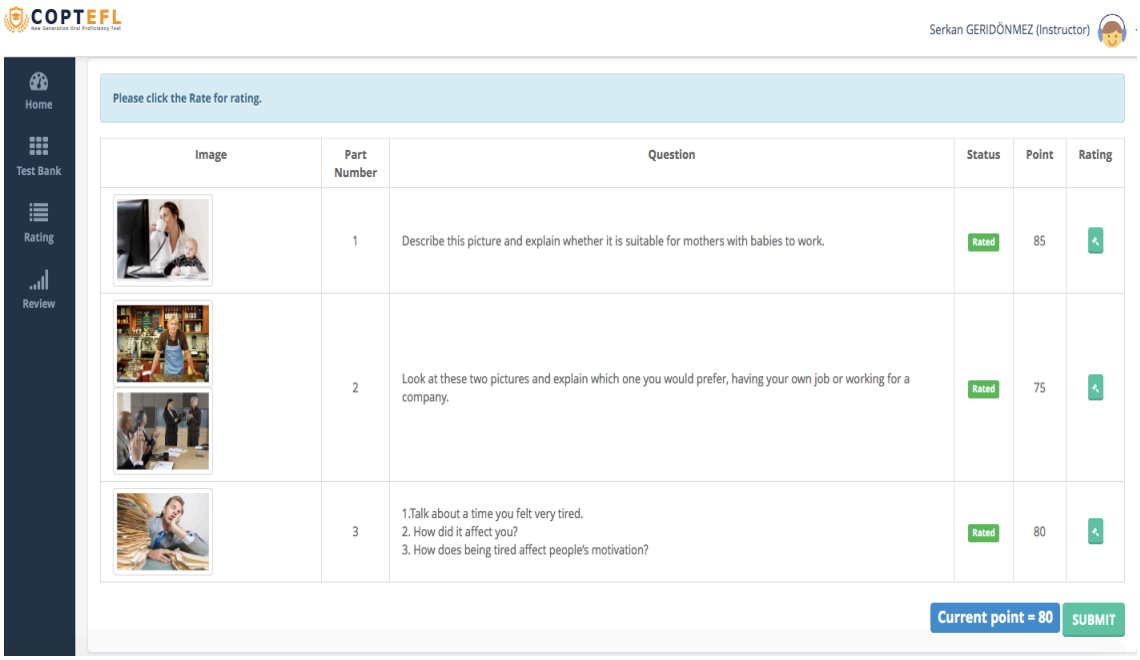



**Figure 3.9.** *Rater module: Completing marking*

**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 account and see the grades they get for each criterion and their final grade on this module (see Figure 3.10).

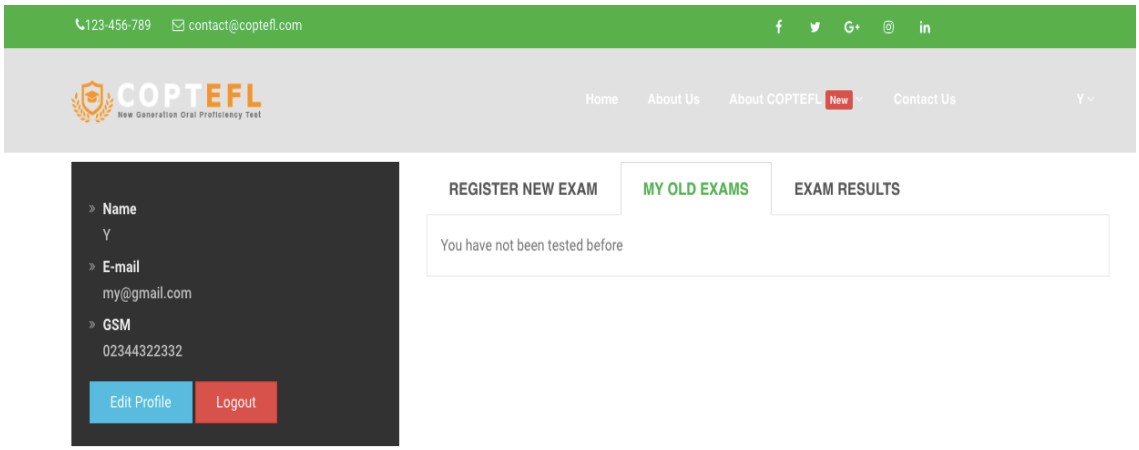



**Figure 3.10.** *Student module*

### 3.5.1.2. *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).

52 Expressing opinions and attitudes using a range of basic expressions and sentences.

53 Comparing and contrast alternatives about what to do, where to go, etc.

59 Describing objects, possessions and products in detail, including their characteristics and special features.

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 the discourse type tasks (see Table 2.1). 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.

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

The tasks and the instructions used in the face-to-face speaking test were the same with the COPTEFL (for an example task for the face-to-face speaking test see Figure 3.11).

**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.

**Figure 3.11.** *An example for the Task 2*

In both tests, the same item pool was used. The aim in making both tests similar in terms of the content of the test was to be able to compare them with each other.

### 3.5.3. Open-ended questions

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 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?

## 3.6. Data Collection Procedure

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

In this part, what processes are followed in the development of the tests in the study was explained. Test development is the entire process of creating and using a test, and having the results of their use (Bachman & Palmer, 1996). The purpose of this section was to explain each step in the process of the test development. Demonstrating how each step was carried out would be useful for two reasons (Bachman & Palmer, 1996). First, it would help increase the accountability of the study; "the ability to say what was done and why" (p.86). Second, it would make it easier to present evidence that "the test was prepared carefully and with forethought" (p.86).

In an effort to adapt the best practices in language test development in the present study, Bachman and 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:

(1) Design,
(2) Operationalization
(3) Test administration.

The following figure presented the test development processes followed in the study (see Figure 3.12).

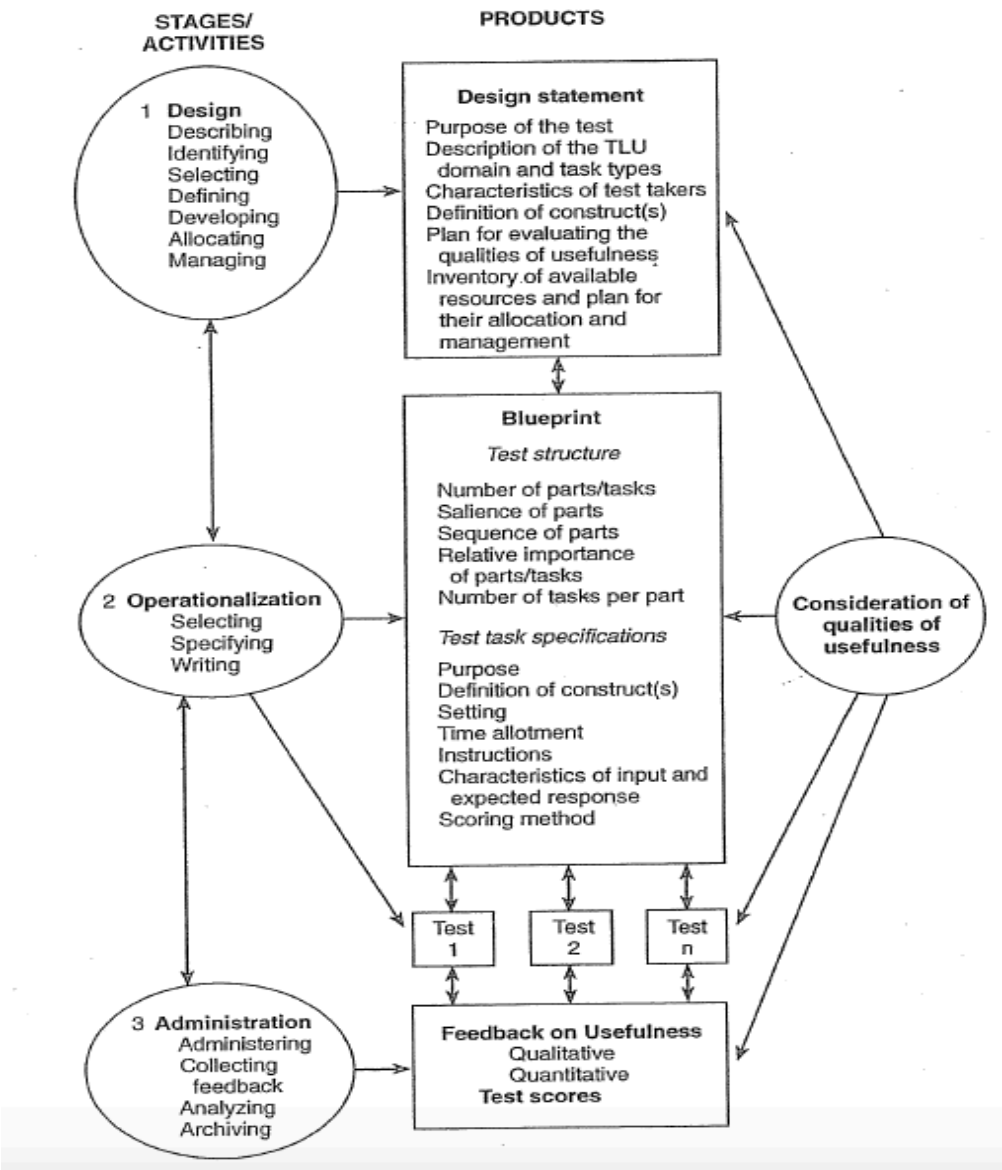



**Figure 3.12.** *Test development process (Bachman & Palmer, 1996)*

### 3.6.1.1. *Stage 1: Design*

Design stage offers a set of principled bases for developing test tasks, a blueprint, and the test, which in turn helps to monitor the subsequent stages of test development (Bachman & Palmer, 1996). There are six activities involved in a design stage (Bachman & Palmer, 1996):

1. A description of the purpose(s) of the test,
2. A description of the target language use (TLU) domain and task types,
3. A description of the test takers for whom the test is intended,
4. A definition of the construct(s) to be measured,

5. A plan for evaluating the qualities of usefulness,
6. Identifying resources and developing a plan for their allocation and management.

The present study followed these six activities. Table 3.1 demonstrated how these activities were carried out in the design stage of the study.

**Table 3. 1.** *Six activities used in the design stage of the study*

| Six activities in a design stage | How were these six activities carried out in the design stage of the study? |
|---|---|
| **1.** A description of the purpose(s) of the test. | **1.** This current study was an English proficiency test in the sense that it was designed to measure the test-takers' ability in spoken English regardless of any training they may have had. The content of this test was not based on any particular language program or course objectives. Unlike an achievement test, the tasks in this test were not based on any pre-specified course content or material. |
| **2.** A description of the target language use (TLU) domain and task types. | **2.** For the TLU domain, see section 3.3. |
| **3.** A description of the test takers for whom the test is intended. | **3.** For the characteristics of the test-takers, see section 3.4. |
| **4.** A definition of the construct(s) to be measured. | **4.** In this study, the Communicative Language Ability (CLA) framework proposed by Bachman (1990) was used as a guide in defining the construct of 'language proficiency' (see Section 2.3). |
| **5.** A plan for evaluating the qualities of usefulness. | **5.** The score comparability findings and test takers' attitudes towards the COPTEFL help ensuring the usefulness of the COPTEFL. |
| **6.** Identifying resources and developing a plan for their allocation and management | **6.** For the resources, see section 3.4.1. |

#### 3.6.1.2. *Stage 2: Operationalization*

In Bachman and Palmer (1996), operationalization stage involves (1) developing test task specifications, (2) a blueprint, (3) writing test tasks and instructions, and (4)

specifying the procedures for scoring the test. This study followed these activities with slight modifications. In the present study, operationalization stage was organized into two sub-stages as (1) test construction and (2) software development. Each has its own activity as illustrated in the Figure 3.13.

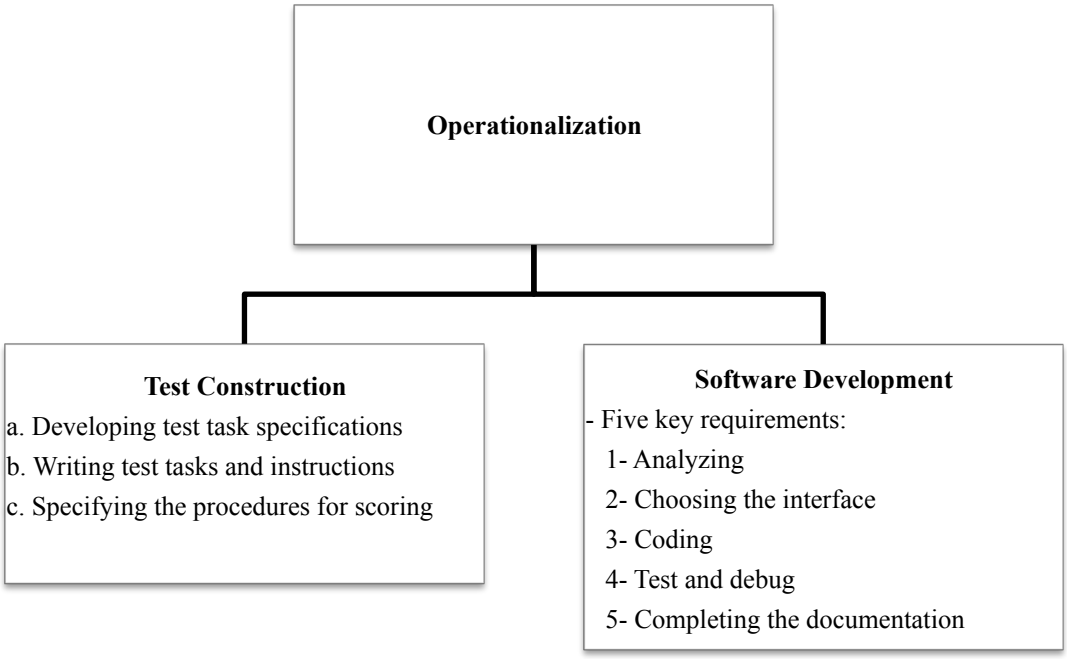



**Figure 3.13.** *Stage 2: Operationalization*

### (1) Test construction

In the present study, test construction stage involved developing test specifications for the types of test tasks –test task specifications and developing the actual test tasks. Test construction also involved specifying the procedures for scoring the test.

In this part, the aim was to discuss what was involved in the process of constructing test task specifications, test items and rating scales. It was intended to detail how the decisions were made during the design process, what the reasons were for those decisions, how the test and the rating scale were then developed. The aim in this documentation was to contribute to a 'validity argument' encompassing all types of evidence that impacts on our understanding of the process at arriving the test construction (Fulcher, 2014). The types of activities that the 'evidence of careful test construction' requires (p.117) were adapted for the study and explained in the sections below.

**a. Developing test task specifications:** Developing the test began with the test task specifications for the various test task types to be included. They are used to permit the development of other tests or parallel forms of the test with the same characteristics. They are the detailed documents that provide the basis for what the test tests and how it tests it. These are needed by test writers and editors who evaluate whether the test has met its aim. The specifications also need to be available to test validators for the test's construct validity.

The purpose for which the test would be used and the target population for the test were our starting point in designing test specifications (see Section 3.3). Having determined these, 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 (see Appendix C). Since the principal user is probably the test writers, the team of test writers were 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 on 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 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 selection of task types, writing of task items, consulting with experts, and pilot tests (see Chapter 4). 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 skill were made at the school. 6 volunteer preparatory programme 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 for the responses. Therefore, the necessary changes and modifications were made after discussing problematic items with the team of test writers (see Appendix D). According to this revision, test task specifications were redesigned to generate test tasks (see Appendix D).

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. It may seem that writing speaking tasks seems relatively painless process compared to writing multiple-choice questions (Luoma, 2004). On appearance, it may seem as if all one needs to do is to write about a topic and leave the test taker to speak (Luoma, 2004). Yet this was not the case in the study. Not all topics lend themselves to item development. The writers needed to find suitable communication activities for the tasks such as expressing an opinion on an issue, defending 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. As Luoma states even with the guide of the test specifications, task writing is a craft (2004). Creativity, sensitivity, insight and imagination are the qualities that underlie much of the success of prospective item writers (Alderson, Clapham and Wall, 1995).

After completing 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 (see Appendix E). 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 (see Appendix G) and shared with the team of test writers.

**c. Specifying the procedures for scoring :** The purpose of the test and the definition of the construct to be assessed were the starting point in selecting the type of scale. In a test of proficiency, it was aimed to design a scale to guide the rating process by focusing on the quality of the performance expected. It was intended to specify a number of particular features for the performances that can be scored separately for each task. In the development of rating scale, an existing scale -which was developed by a formal testing body in the institution, 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 rating scale evolved (see Appendix B). 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 developed, 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.

**(2) Software development**

***Programming language***

The programming language of the COPTEFL system is C#. Microsoft created C#, a new programming language based on the C and C++ languages, but it was developed from the ground up. Microsoft started with what worked well in C and C++ and included new features that would make these languages easier to use. Microsoft describes C# in this way (Chand & Gold, 2002):

> "C# is a simple, modern, object–oriented, and type safe programming language derived from C and C++. C# is firmly planted in the C and C++ family tree of languages and will immediately be familiar to C and C++ programmers. C# aims to combine the high productivity of visual basic and raw power of C++" (p.1).

C# has been developed according to the current trend and is very powerful and simple for building robust applications. It involves built in support to help any component turn into a web service that can be invoked over the Internet from any application running on any platform.

***Programming stages***

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 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 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 analysing phase. In order to translate requirements into design, meetings between the researcher and the software developer were held twice in 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 sample 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 documents bugs if there were any. The developer did the necessary changes on 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. These were stated in section 4.2.3.1.

#### 3.6.1.3. *Stage 3: Test administration*

This stage involves giving the test to a group of test takers, collecting data and analyzing data (Bachman & Palmer, 1996). It has two phases as try-out and operational testing (Bachman & Palmer, 1996). The aim in the try-out phase is to gather information on the usefulness of the test itself and for the improvement of the test and testing procedures. Revisions are made with regard to the feedback obtained from a tryout. Operational testing has also the aim to gather information on the usefulness of the test, but this time administering the test involves the goal to accomplish the specified use/purpose of the test. The present study involved two phases as (1) try-out for the COPTEFL system and test tasks, (2) operational testing for the COPTEFL and the face-to-face speaking test (see Figure 3.14).

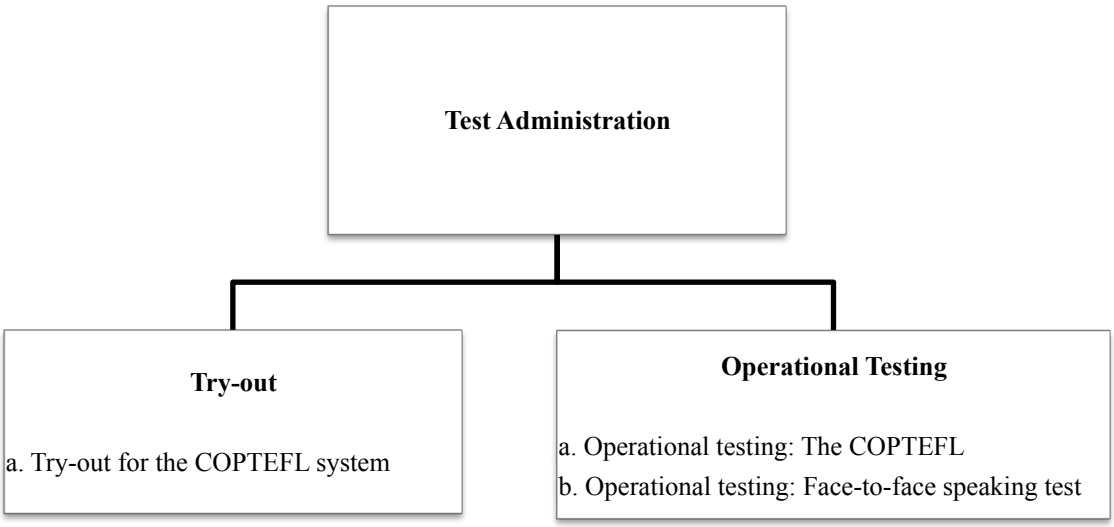



**Figure 3.14.** *Stage 3: Test administration*

### (1) Try-out

**a. Try-out for the COPTEFL system:** After the development of the software, the next step was to test it with users. These were the students having their regular laboratory classes as part of their language program. After getting the necessary permission from the class teacher, students were informed that the goal was to test the system, see whether it would work as it was planned and they would not be graded. So, it was announced that although they performed the test, they would not be given grades. 15 students who consented to participate were included in the trial.

The researcher and one of the software developers conducted the test. The software developer analyzed the performance of the software and took notes on additional requirements. The researcher gave the instructions for the test and worked with the software developer in calculating the strengths and weaknesses of the test system by observing each test taker's test screen. The results of the trialing were as follows:

1. It was very important for the test system to be able to record non-stop as test takers talk, upload the captured audio to a server via browser, and save voice data and play back. During setting up and testing microphones, the system received error in accessing the microphones. The computers had support wizard for troubleshooting

microphone problems. The support audio wizard set the default microphone. This adjustment required removing and then plugging the microphones into the port on the computer. The problem was fixed after following those steps.

2. The system showed the prompt with "Allow" or "Disallow" permission to microphone. Students were told to click on "Always allow" to give permission. After changing dropdown menu to "Always allow", it was programmed not to ask for permission during examination. But, for each task, it required permission to allow and that extension for permission did not disappear after clicking on. This interrupted the flow of the testing process. Another problem related to giving permission for microphone was that in each click, the sound recording stopped and started again. The software developer took notes on these problems to do debugging.
3. There was software – coding error that caused to produce incorrect question ordering. Questions for part 3 were the same as Part 2. This was one of the serious problems experienced during the trial process.
4. Although the test system had some bugs to fix, it was satisfactory that the system captured the sounds and recorded the tracks of each test taker's sound simultaneously to the server.

The first step before the COPTEFL system was used for a "live" administration was to conduct trialing. In cases where new procedures are being tried out, such piloting is essential. Without it, problems could not become apparent until it was too late. Once the trial was completed, the results were taken into account and all the necessary adaptions were made regarding the test, the procedures or the computer system. After trying the adapted system with various other users, the system was ready for a "live" administration where the test is being used for its intended purpose. This required a detailed planning for a smooth test administration.

**(2) Operational testing**

**a. The COPTEFL:** After the trialing sessions were completed and necessary changes were made due to the results, a "live" administration of the test was done. One weekbefore 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. A total of 45 volunteer preparatory program students from various proficiency levels participated in the administration part of both tests. 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.net

**STEP 2:** Test takers entered the testing system. They entered their personal information. Once the test takers typed their personal information such as name, surname and personal identification number, a unique ID was automatically generated for each test-taker. This ID included the test takers' name and a numerical code for privacy purposes and stored in an underlying location in server database.

**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. In order to take into account the validity of the procedure, these instructions were given in a simplified language to ensure that the wording was equally understandable to all test-takers of different ability levels.

**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 expresses 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' the attitudes towards the test modes. The questionnaire collected data in two areas: (a) general attitudes towards the COPTEFL system and (b) a direct comparison of two delivery modes. Students were asked to write down their opinions in relation to both areas. After they completed writing, the researcher collected the answer sheets.

Test taker consent forms were provided to have permission from the test takers for their results to be used. The results were reported to test takers with their score band for each area assessed by the test. Recordings of the tests were decided to be stored for a year in case there are questions or complain about results. The results of the administration experiences are as follows:

1. It is important to plan the registration process for computerized tests. Although test takers had registered in advance, some of them did not show up. Some of other students who did not pre-register wanted to participate. Since the tests were constructed for each student on the admin panel, at least one computer must be reserved for the researcher in order to make necessary changes during the test.
2. Before the test begins, proctors should be informed with seating arrangement for the lab. They need to be told the number of the seats that should be left empty between test takers. In the exam, proctors had to change their shift. Since the new comer had no idea about the seating plan, the test takers were seated directly behind each other. This made test takers have trouble in concentrating on their own test.

**b. Face-to-face speaking test:** Standardized test booklets were designed for each interviewer to be administered under consistent procedures so that the test-taking experience was as similar as possible across test takers and interviewers. In the booklet, the steps for the test procedure were explained and the instructions that the instructors should read to the test takers were provided along with the test. The test format was the same with the format in the COPTEFL system. Each student had different test and also the students in the same sessions had different test.

Once the administration of the face-to-face speaking test completed, interviewers reported their views on the examination and explained which worked well, and which not. The results of the trialing based on the views of the teachers are as follows:

1. All of the interviewers reported that the standardization for the administration of the exam increased the likelihood of valid and reliable assessment. They emphasized that the similar experience would increase the fairness of the test as well as making test takers' scores more directly comparable with each other.
2. One very important thing to do is making a plan for interviewer norming with samples from the test. In the face-to-face speaking test, it was observed that some teachers let the students read the questions on their own while some others read themselves aloud. Although interviewers were told to read aloud, they got confused when they saw questions written on the pictures given students.
3. Thinking and response time for each task made it problematic for interviewers to keep track of time since they did not have countdown timer.
4. Students were not used to time restriction for tasks in a face-to face speaking test. They started to speak without waiting for thinking time. Similarly, they had problems in using the given time; they talked either more or less than necessary.
5. Part 1 in the test was not fully answered by some of the students. In this part, students were asked to describe a picture and give an opinion about a topic related to the picture. They forgot to describe the picture. In relation to this problem, some interviewers suggested to divide this question into two parts as description and giving opinion.
6. Part 3 in the test was not fully answered. In this part, students were given three related questions and asked to discuss about them by covering each. Some of them forgot answering each.

### 3.7. Data Analysis

#### 3.7.1. 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, 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. Howell (2002) indicates that there are different approaches in estimating inter-rater reliability (i.e. calculating the percentage of times that two judges agree on their rating or correlating the ratings of two judges), but maybe the best way is looking at the

ICC. The reason is that ICC does not only take into account the correlation between raters, but also examines whether the actual scores they gave test takers differed.

Before proceeding to comparing 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, 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 ICC.

### 3.7.2. Test-taker attitudes

The open-ended questions exploring the test-takers' attitudes towards the test delivery modes were analyzed based on the qualitative content analysis scheme of Creswell (2012). 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 English Language Teaching Department as a research assistant and has experience in qualitative data analysis analysed 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 researches, a final consensus on the categories was achieved.

# 4. RESULTS

## 4.1. Introduction

This chapter reported on the aspects of a priori validation of the test tasks used in the study in terms of their cognitive demands and contextual features and a posteriori validation through a discussion of scoring validity. The first and second research questions, which are about cognitive and context validity, were investigated based on the task characteristics. The third research question, which relate to scoring validity, was examined on the statistical analyses of the raw scores assigned to the test takers. The last research question was explored through the qualitative content analysis scheme of Creswell (2012).

## 4.2. Investigation of Theory-based Validity

The first research question sought evidence to support the interpretations made about theory-based or cognitive validity. For this purpose, the following research question was posed:

Research Question 1: What are the cognitive demands of the speaking test tasks?

For the investigation of the first research question, the test specifications for each test task were constructed regarding Weir (2005) and the GSE descriptors for speaking skill. The results for each task were presented below.

### 4.2.1. Warm-up

Warm-up task in the study was designed to put test takers at ease, and to allow them to make the transition to speaking in the target language and to become comfortable with the test situation. For this purpose, the task was designed as the easiest task in the test. Test takers respond to a question on a personal topic “Please introduce yourself”. The task aims at eliciting short responses from direct, personal knowledge and experience, and therefore requires lower level cognitive demands. Test takers are given 30 seconds to reply. Since the aim is to provide reassurance to test takers to reduce their anxiety levels, no scoring takes place for this task. A screen shot from the COPTEFL exam page was presented in Figure 4.1 below.

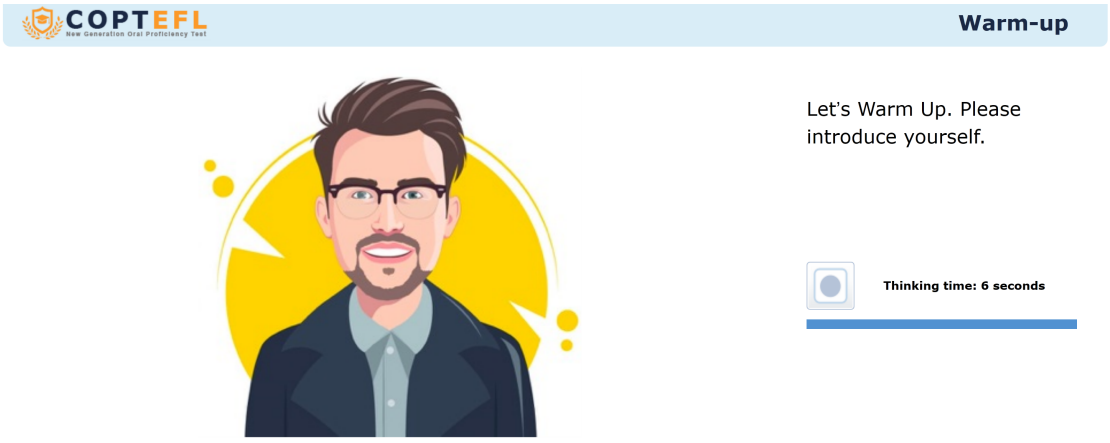



**Figure 4.1.** *Warm-up part from the COPTEFL*

### 4.2.2. Speaking task 1: 51–58/B1 (+) level

**(1) Conceptualization:** Task 1 requires test takers to respond to two questions related to one picture prompt. Test takers are asked to first describe what they see on the picture given. In this picture, they see a person or people engaged in a concrete, everyday activity. Then, they are asked to elaborate by talking about the same topic given in the picture in more general terms and provide an opinion with reasons and justification (see Figure 4.2).


**Figure 4.2.** *Task 1 from the COPTEFL*

Although the context in this task is concrete, it involves fairly abstract concepts based on this concrete context. This part of the question is conceptually more

demanding compared to first part in which the aim is simple description. The Task 1, therefore, requires non-personal information and more open-ended responses.

**(2) The support provided:** In this task, test takers are given a visual (non-verbal) prompt which may give them ideas that they may wish to express. In both tests –the COPTEFL and the face-to-face speaking test, the rubric is presented with oral and written prompt. The nature of the input provided is pre-recorded for the COPTEFL and scripted for the face-to-face speaking test. The nature of the interaction is the computer-test taker for the COPTEFL and interviewer-test taker for the face-to-face speaking test.

**(3) Grammatical encoding:** Table 4.1 demonstrates the functions that are required from the test takers in task 1. They are based on an investigation of GSE descriptors for the level of 51–58/B1 (+).

**Table 4.1.** *Functional requirements of the task 1*

| Task No | Functional Demands |
|---|---|
| Task 1 | Describing<br>Elaborating<br>Explaining opinions<br>Justifying opinions |

As can be seen in the table 4.1, the task requires multiple functions. However, these functions are required to perform in a step by step progression. Test takers are firstly asked to describe what is happening in a photo and then elaborate/explain or justify on their opinions.

**(4) Phonological encoding:** Phonological encoding is utilized in the present test in the Fluency component of the scale. Below are sample scale descriptors of fluency (see Table 4.2).

**Table 4.2.** *Descriptors of fluency*

| Assessment component | Descriptors |
|---|---|
| Fluency | Speech is effortless with speed that approaches that of a native speaker.<br>Speaks smoothly with little hesitation.<br>Speech is slow and often hesitant and jerky. Sentences may be left uncompleted, but speaker is able to continue, however haltingly.<br>Noticeable pausing, false starts, reformulations and repetition. Difficult for a listener to perceive continuity in utterances. |

**(5) Time pressure and reciprocity conditions:** Task 1 requires test takers to give two responses when they are asked to give answer. This may place cognitive demands on the test taker in this respect. So, they are given 15 seconds for planning their answers and 90 seconds to give a response.

### 4.2.3. Speaking task 2: 51-58/ B1(+) level

**(1) Conceptualization:** The focus in this task is on comparing and contrasting and expressing a preference in relation to the aspects already elaborated. Test takers are given two pictures which provide the basis for contrast and comparison. They are required to express why they would prefer a certain aspect or idea illustrated through the pictures over the other aspect or idea. A screen shot from the face-to-face speaking test booklet was presented in Figure 4.3 below.

**Part 2:** Look at these two pictures and explain whether you would prefer spending your childhood in the past or future.

**15 sec. for preparation** **90 sec. for reply**

**Figure 4.3.** *Task 2 from the face-to-face speaking test*

In order to show their preferences, test takers need to give brief reasons for advantages and disadvantages and explain the main points in an idea or problem with reasonable precision. Although the task is regarded to impose lower cognitive requirements with respect to the nature of the information required, the rhetorical functions required by the question i.e. speculating, justifying viewpoints can have higher cognitive demands.

**(2) The support provided:** The tasks in the study differ in the number of visual prompts they provide. Two pictures are used for Task 2 and they need to be compared and contrasted, and make higher lexical requirements. The conditions for the presentation of the rubric, the nature of the input and interaction are the same with Task 1.

**(3) Grammatical encoding:** Table 4.3 shows the functions that are required from the test takers in task 2. They are based on an investigation of GSE descriptors for the level of 51–58/B1 (+).

**Table 4.3.** *Functional requirements of the task 2*

| Task No | Functional Demands |
|---|---|
| Task 2 | |
| | Describing<br>Comparing<br>Elaborating<br>Expressing preferences<br>Explaining opinions<br>Justifying opinions<br>Summarizing |

As can be seen in the table 4.3, the task requires multiple functions. As in the Task 1, these functions are required to perform in a step by step progression. Test takers are firstly asked to compare some aspects of the photos and provide an opinion and express a preference in relation to the aspects already elaborated. Although the Task 2 targets the same GSE range (51–58) with the Task 1, the functions operationalized in this task differ from Task 1.

**(4) Phonological encoding***:* Phonological encoding is operationalized in the present test in the Fluency component of the scale (see Table 4.2).

**(5) Time pressure and reciprocity conditions:** In the Task 2, test takers have 15 seconds for preparation and 90 seconds for giving their responses.

### 4.2.4. Speaking task 3: 59–66/B2 level

**(1) Conceptualization:** Task 3 is considered to be the most conceptually challenging task in the test. In this task, test takers are given three related questions. In the first question, they are asked for a description of personal experience in relation to an abstract topic. In the second question, they are asked to elaborate on the same topic regarding their feelings or attitudes. In the third one, they are asked for a more objective

discussion of the topic from the perspective of wider relevance to society/people in general (see Figure 4.4).

**Part 3:** 1. Talk about the last time you cried.
2. How did you feel after it?
3. Do you think crying makes a person feel relieved?

**1 min. for preparation** **3 min. for reply**

**Figure 4.4.** *Task 3 from the face-to-face speaking test*

During one-minute preparation time, test takers are required to integrate their ideas regarding three questions into a long turn. For three minutes, they are required to give opinions and justify their opinions. This task requires test takers to go beyond description to argumentation and therefore, the need for conceptualization is considered to be high.

**(2) The support provided:** A picture is given for extra contextualization but is not referred to in the task.

**(3) Grammatical encoding***:* Table 4.4 presents the functions that are required from the test takers in task 3. They are based on an investigation of GSE descriptors for the level of 59–66/B2 level.

**Table 4.4.** *Functional requirements of the task 3*

| Task No | Functional Demands |
|---|---|
| Task 3 | Providing personal information<br>Describing<br>Elaborating<br>Explaining opinions/preferences |

| | |
|---|---|
| | Justifying opinions<br>Summarizing |

As different from the Task 1 and 2, in this task, test takers are required to integrate their ideas. So, they do not necessarily use the grammatical functions in a step by step progression. Rather, they are required to provide a coherent and cohesive long turn.

**(4) Phonological encoding:** Phonological encoding is operationalized in the present test in the Fluency component of the scale (see Table 4.2).

**(5) Time pressure and reciprocity conditions:** In the Task 3, test takers have one minute for preparation and three minutes for giving their responses.

## 4.3. Investigation of Context-based Validity

This section addressed the second research question:

Research Question 2: What are the contextual characteristics of the speaking test tasks?

The contextual parameters of the tasks were investigated based on the aspects of context validity for speaking proposed by Weir (2005).

### 4.3.1. Test setting

#### 4.3.1.1. *Response format*

Response formats used in the present tests are short responses, extended responses and individual long turn. By including different types of response format, it was aimed to bring variety and provide opportunity for producing a wider range of speech functions. Below are response formats for each task used in the tests (see Table 4.5).

**Table 4.5.** *Response formats used in the tests*

| Task No | Task Formats |
|---|---|
| Warm-up | Short response |
| Task 1 | Extended responses |
| Task 2 | Extended responses |
| Task 3 | Long turn |

#### 4.3.1.2. *Task purpose*

In the tests, the purposes of the tasks were introduced to the test takers with oral and written prompts. For the Task 1, before moving on to the response, it is explained that "I will ask you to describe a picture and answer the question related to the picture You will have 15 seconds to think about your answer and 90 seconds to reply." This task taps into several functions: the test takers are asked to describe, explain and justify their opinions. For the Task 2, it is stated that "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." This task also taps into several functions: the test takers are asked to compare, comment, explain and justify their opinions. Finally, for the last task, Task 3, it is given that "I will ask you three questions in this part. You will have one minute to think about your answers before you start speaking. You will have three minutes to answer all three questions". In this task, the test takers are asked to provide personal information, elaborate on and/or explain their feelings and justify their opinions.

#### 4.3.1.3. *Weighting*

In the tests of the study, each task and assessment dimension are equally weighted in their contribution to the overall score. They are assessed individually and an overall score is awarded to the test takers on the three tasks according to Pronunciation, Fluency, Grammar range and accuracy, Adequacy of vocabulary for purpose and Task fulfillment (see Appendix B).

#### 4.3.1.4. *Known criteria*

This contextual parameter is related to whether the test takers have known the assessment criteria of the test. The test takers in the study were not given information about the assessment criteria before the tests. However, when the scores were announced them, they got aware of the assessment dimensions since the scores were introduced with each assessment dimension.

#### 4.3.1.5. *Order of tasks*

In the tests, the tasks were ordered according to their assumed difficulty based on the GSE scale descriptors. The tasks targeted specific range in the scale. While the Task 1 and Task 2 targeted the same range, the functions allocated for the tasks were

different. Only in the Task 2, the test takers are required to compare and contrast. However, in the Task 1, they were given a picture and asked to describe and provide their opinions. The order of the tasks was presented in Table 4.6.

**Table 4.6.** *The order of the tasks and their relevant GSE levels*

| Order of task | Task No | GSE Level | Task Formats |
|---|---|---|---|
| 1 | Warm-up | 22-42/A1-A2(+) | Short response |
| 2 | Task 1 | 51–58/B1 (+) | Extended responses |
| 3 | Task 2 | 51–58/B1 (+) | Extended responses |
| 4 | Task 3 | 59–66/B2 | Long turn |

#### 4.3.1.6. *Time constraints*

In the tests of the study, each task involves preparation and response time. In the Task 1 and Task 2, the test takers are allowed to use 15 seconds to think about their answers and 90 seconds to give their responses. In the Task 3, they are given one minute to plan their answer and three minutes to give answer since this task is assumed to be cognitively more demanding.

### 4.3.2. Task demands

#### 4.3.2.1. *Channel of communication*

This study includes two testing modes as semi direct and direct testing. In order to make the tests comparable, both tests were designed similar in the way tasks are delivered. That is, although direct testing enables interaction so that communication can be co-constructed by the test takers and interviewers, the interviewers in the present study were not allowed to provide further speech other than the presentation of the rubric and the questions. Although the channel of communication was through interviewer-to-test taker in the face-to-face speaking test, the types of the tasks which were monologic in nature helped us having similar test conditions for the delivery of the test in terms of channel of communication.

#### 4.3.2.2. *Discourse mode*

Discourse mode was operationalized in the test tasks in relation to the functions identified in the GSE (GSE, 2015). For a discussion of the functions, see the sections on 4.2.

#### 4.3.2.3. *Input length*

Since the input length has an impact on cognitive complexity, it was aimed in the study to provide short prompts which do not impose heavier cognitive requirements. So, the sentences in the rubric and instructions were kept simple and short.

#### 4.3.2.4. *Nature of information*

The nature of the information in the present tests progresses from concrete to fairly abstract. In the Task 1, the test takers are required to talk about concrete information first (picture description) and then they are asked to give/express/justify an opinion which is fairly abstract since it involves eliciting ideas. In the Task 2, test takers are asked to choose one picture and compare the aspects revealed in the question. So, this task requires test takers to hypothesize about the situations or conditions given in the pictures and due to their hypotheses, they are required to come up with a choice in relation to their preferences. Since the test takers are required to go beyond concrete content to hypothesize about what would be their choice, this task is considered to be fairly abstract. In the Task 3, questions include both personal and non-personal information. While the first and second questions ask for details related to test takers personal experiences, the third question asks for a discussion of the topic from the perspective of wider relevance to society or people in general. So, this task also can be considered as fairly abstract.

#### 4.3.2.5. *Topic familiarity/content knowledge required*

The topics provided in the tasks are given below:

- Daily life
- Descriptions of buildings/places/people
- Education/ school life
- Environmental issues
- Family
- Food and drink
- Health
- Entertainment
- Relationships
- Science

- Technology
- Travel and tourism
- Work and job related

#### 4.3.2.6. *Lexical resources, structural resources, and functional resources*

Demonstration of resources is at the specific GSE level that each task targeted. For the Task 1 and Task 2, demonstration of grammatical control at the 51-58/B1 (+) level necessary for a rating of 3 (out of 5) for the task. 51-58/B1(+) lexis is sufficient to respond adequately to the task. As for the Task 3, demonstration of grammatical control at the 59-66/B2 level necessary for a rating of 3 (out of 5) for the task. 59-66/B2 lexis sufficient to respond adequately to all questions. For the functional recourses, see section on 4.2.

## 4.4. Score Comparability

Research Question 3: 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?

This section presented the results of quantitative analyses of test takers' performances in the COPTEFL and the face-to-face speaking test. It was composed of three parts. The first part evaluated the consistency of ratings. The second part examined the order effect on test scores. The final part described the comparability of raw scores across the two modes.

### 4.4.1. Inter-rater reliability

In speaking tests where the scores were gathered from two raters, a type of error variance arising from the individual raters involved in the process. To what extent they were in overall agreement needs to be estimated. Thus, inter-rater reliabilities were calculated to estimate the consistency of marking between raters for each test delivery mode.

In order to investigate the inter-rater reliability, two-way random absolute agreement intra-class correlation coefficient (ICC) was run (see Table 4.7).

ICC results for overall test scores and each task type test scores across delivery modes were provided in the tables below.

**Table 4.7.** *Intra-class correlation coefficient (ICC) results for inter-rater reliabilities across test delivery modes*

| Test score | COPTEFL ICC | Face-to-Face ICC |
|---|---|---|
| Overall score | .806* | .889* |
| * – significant at the .01 level | | |

Table 4.7 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 .9 indicated a good reliability for both tests.

**Table 4.8.** *Intra-class correlation coefficient (ICC) results of inter-rater reliabilities for each task type across test delivery modes*

| Task type | COPTEFL ICC | Face-to-face ICC |
|---|---|---|
| Opinion | .686* | .796* |
| Comparison | .702* | .842* |
| Discussion | .772* | .890* |
| * – significant at the .01 level | | |

Table 4.8 revealed the ICC results for each task type. 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 .5 and .75, meaning that the level of reliability was moderate. For discussion task, it was .77, which indicated a 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.

### 4.4.2. Order-effect

Prior to comparing raw scores awarded to the two delivery modes, a one-way Analysis of Covariance (ANCOVA) was computed to assess whether the existence of order has an effect on speaking test scores. Dependent variable was the test scores while the independent variable was the type of tests. The findings revealed that there is a significant interaction of test order and test scores ($F(1,43)=6.89$, $p=.012$). As Figure 4.1 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.

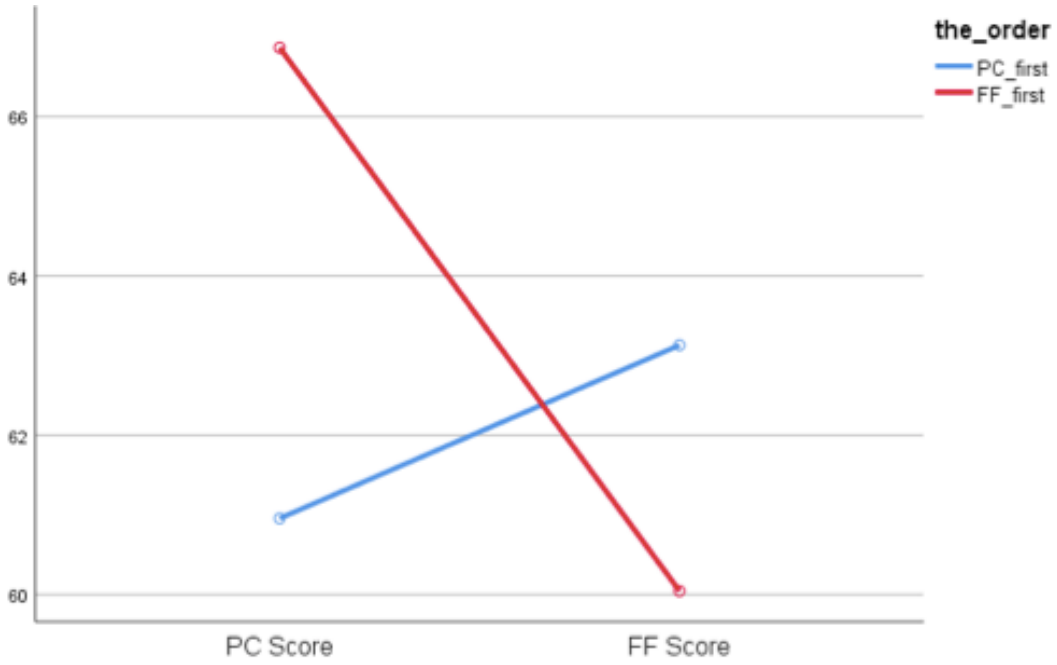



**Figure 4.5.** *The interaction between group and test mode*

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.

### 4.4.3. 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 4.9)

**Table 4.9.** *ICC results for total scores and task type scores across test delivery modes*

| | | COPTEFL | | |
|---|---|---|---|---|
| | Total score | Opinion task | Comparison task | Discussion task |
| Face-to-face | | | | |
| Total score | .632* | | | |
| Opinion task | - | .382 | | |
| Comparison task | - | - | .576** | |
| Discussion task | - | - | - | .704* |

* – significant at the .01 level
** – significant at the .05 level

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.

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 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 4.10).

**Table 4.10.** *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 4.8 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.

### 4.5. Test Taker Attitudes towards the Test Delivery Modes

Research Question 3: What are the attitudes of test takers in relation to the test delivery modes?

In order to explore test takers' the attitudes towards the test delivery modes, open-ended questions were given immediately after the administration of the test. The questionnaire collected data in two areas: (a) general attitudes towards the COPTEFL and (b) a direct comparison of two delivery modes. In relation to the general attitudes towards the COPTEFL, three categories were generated from the data. As for the direct comparison of two modes, three categories were generated for each delivery mode.

***a) 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 4.11).

**Table 4.11.** *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 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

**1. Test system:**

Nearly all of the participants stated that the test system was well designed, quite easy to operate, and works well. 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. Although most of the comments were positive in relation to the test system, there were a few constructive comments for improvement of the system. In those comments, test takers stated that microphones

could be better in order to achieve best possible results for sound recording. Also, some of them referred to the bad effects of seeing countdown timer on the screen. Apart from those, one of the participants also made a comment on transition from instructions to tasks. According to her, that transition was abrupt and due to this, she got anxious.

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

*"Visuals and the system interface were motivating."*

*"The COPTEFL system is well designed that anyone can use it."*

*"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."*

*"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."*

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

**2. Tasks:**

Some of the test takers remarked on the difficulty level of the tasks. While many of them judged the tasks at the right level of difficulty, one of the test takers stated that tasks were tough to answer.

*"Difficulty of the tasks were at the right level for me."*

*"Tasks were easy to do."*

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

**3. Time limit:**

Most of the test takers reported that the time allocated for thinking and response was satisfactory for them. But, some of the test tsakers either found thinking time or response time little and recommended to multiply it.

*"Thinking time made me better concentrate on my answers and therefore I performed better."*

*"Thinking and responding time were enough for me."*

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

*"Thinking time was a little bit short for me."*

### *b) 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 favored the face-to-face mode. These were presented in order of frequency below (Table 4.12).

**Table 4.12.** *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

**1. Interaction:**

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."*
>
> *"When there is a person, it is more relaxing to be examined. That is why I prefer face-to-face speaking exams."*
>
> *"Interaction with the teacher helped me speak more."*
>
> *"When you have a burden ahead of you or when you get anxious during examination, there is nobody to accommodate your answers in computerized tests. But, in face-to-face speaking exams, interaction with the interviewer can prevent that possibility of feeling panic."*

**2. Naturalness:**

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 face-to-face speaking test as a better measure of their spoken English because of this sense of naturalness.

*"Thinking time and having no one caused me feel that computerized exams are not testing naturally."*

**3. More time:**

A few of the test takers stated that they distracted by the countdown timer on the screen of the computer while they were speaking and they had difficulty to finish their responses. However, having an interviewer made them comfortable to use the necessary time to finish their last sentences.

*"It was difficult to match the end of your sentences with the end of seconds on the screen. In the face-to-face speaking test, you can finish your last words."*

Although some of the test takers preferred having 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.

The following categories were generated from the responses of those who preferred the COPTEFL.

**1. Less anxiety:**

Test takers reported a lower level of anxiety in the computer mode than face-to-face mode. They attributed the reduced level of anxiety to the lack of interaction with an interviewer. They felt talking to a specific teacher who was perceived as discouraging would be more anxiety provoking. When talking alone, on the other hand, they felt like they were not taking a test. Having an interviewer also made them over concerned about the feedback as a sign for good performance. When they did not receive some warm feedback, they felt nervous due to the possibility of making a mistake. In computer mode, they did not have such concerns since they had already known that there would be no reactions.

*"The COPTEFL was much more relaxing compared to face to face speaking exams."*

*"The COPTEFL was more effective in testing and it made me feel comfortable during the examination."*

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

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

*"One of the best advantageous of the COPTEFL is that you got free from any of those specific teachers who has a potential to make you feel irritated in the test."*

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

**2. Better control:**

Preparation and response time caused the test takers to have better control of the pace of test taking. This, in turn, helped them to better concentrate on thinking and responding.

*"When compared to the face-to-face speaking test, the COPTEFL provided more control in using thinking and responding time. You see how longer you have to talk on the screen."*

**3. Test fairness:**

Some of the test takers stated that no influence of the interviewer made the COPTEFL more valid since it provided opportunity to have the same administration under the same conditions.

*"The COPTEFL is much more objective than the other speaking tests due to the same administration conditions and controlled processes by the experts before the test."*

*"This test is much more valid than face-to-face speaking tests in terms of evaluation."*

*"Different interviewers have different styles during the tests. If her/his style does not help you understand, you would have difficulty. But, in the computerized test, you don't have such a problem."*

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.

Apart from reporting on attitudes towards both test delivery modes, some of the test takers provided their suggestions on how to improve the COPTEFL system. In the Table 4.13 below, those suggestions were summarized.

**Table 4.13.** *Summary of the suggestions on the COPTEFL*

| Suggestion Topics | Suggestions |
|---|---|
| **1. Test System** | |
| | Start responding when you are ready |
| | Finish responding when you complete |
| | Add "ask again" button |
| **2. Tasks** | |
| | Bring variety for tasks types |
| | Bring variety for pictures |
| | Address current issues in tasks |
| | Include starting questions |
| **3. Time limit** | |
| | Use bar chart instead of countdown timer |
| | Use no sign for time limit |
| **4. Administration** | |
| | Provide COPTEFL for laboratory classes |
| | Allow using COPTEFL at home |
| | Administer to one student at a time |

Students' suggestions were divided into four categories as test system, tasks, time limit and administration. In relation to the test system, they wanted to have more control for starting and finishing speaking. One of them asked for "ask again" button to have after they completed answering. As for the tasks, they wanted us to bring some variety with regards to task types, content of the tasks and pictures. Also, one of them stated that it would be better to have more starting questions at the beginning of the test which could make them better warm. For the time limit, it was seen that countdown timer put the pressure on test takers and some of them preferred bar chart instead of a countdown timer while one of them preferred nothing to see on the screen. Finally, as for the administration, students wanted us to allow using the COPTEFL at their laboratory classes and at home. Some of them stated their discomfort in relation to administration environment. Since there were other students taking the test at the same time, the noise them tense. Some of the suggestions were provided below.

*"I want to start speaking when I am ready. So, I would like to plat the start button whenever I want."*

*"Ask again button could appear when we finished answering."*

*"There must be starting questions. The exam finished early."*

*"Pictures must be enriched."*

*"Current issues should be asked as well."*

*"Instead of having seen a timer that showing the seconds, it would be better to see a bar chart in which there is no sign for the seconds left."*

*"I want to access to this system at home and take an exam. This system can provide us a chance to practice our spoken English anytime we want."*

*"It would be great to use this exam in our lab classes because the tasks that we do in this lesson were not as effective as this for making us speak."*

*"Hearing other test takers' voices were disturbing."*

Some of these suggestions required us to take some cautions and make some changes. Especially, the influence of the noise was a problem that could affect the reliability of the scores. Noise-canceling microphones can be used next time and the number of students taking the exam can be decreased. Apart from that, a bar chart can be preferred instead of a countdown timer and varieties can be offered for the content of the tasks. Some of the suggestions, however, cannot be taken into consideration. Allowing students to have control over starting and finishing the time when they wish can lead to non-standard administration of a high-stake test where the aim is to decrease the possibility of variances over performances.

# 5. DISCUSSION

## 5.1. Introduction

The present study attempted to document how the Computerized Oral Proficiency Test of English as a Foreign Language, namely the COPTEFL was developed and what processes were followed in validating this newly developed speaking test. From the quantitative research perspective, it was aimed to discover whether the scores obtained from the administration of the COPTEFL was comparable with the conventional face-to-face speaking test. From the qualitative research perspective, the study explored test takers' attitudes towards the two test delivery modes.

## 5.2. Inter-rater Reliability

An important step in developing a test is the statistical analysis of the degree in the consistency of the ratings (Bachman & Palmer, 1996). If the ratings of a test are not reliable then all of the quantitative information obtained to compare the scores will have been a waste of time. So, before presenting the possible reasons that may impact on the results of correlation coefficient analysis, it is important to raise rater reliability matter.

Rater reliability is related with the extent to which two or more raters are able to agree with each other on the test scores they give to the same test taker(s) (Fulcher, 2014). As McNamara (2000) states, introducing raters into assessment procedure is both necessary and problematic. It is problematic due to the subjectivity inherent in spoken language assessment (Szmerdt-Chandler, 2011; McNamara, 2000). That is, the scores given to test takers are not only the reflection of their performance but also the raters' judgments (McNamara, 2000). The subjectivity in the rater judgments is one of the major source of measurement error and a threat to the reliability and validity of test scores (Bachman, Lynch & Mason, 1995). Therefore, the quality of rater scoring decisions has significant consequences for test reliability and validity. A reliability of .80 is the minimally acceptable level in high-stakes testing (Carr, 2011). Inter-rater reliability estimates in the present study was .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. There are various

factors that can interfere with a rater's ability to give consistent judgments. One of them is the problem with the interpretations of rating scales (Bachman & Palmer, 1996; Alderson, Clapham & Wall, 1995). Raters' different interpretations of the rating scales can affect the test scores and therefore, can lead to inconsistent scoring (Fulcher, 2014; Bachman & Palmer, 1996). One reason for different interpretations of scales may be that the terminology of the band descriptors is vague. As Bachman and Palmer (1996) state:

> "If we ask raters not sophisticated in linguistics to rate an oral interview on 'knowledge of register' without first carefully defining what we mean by 'register', some raters might interpret the use of a different regional variety as register. This can be a particular problem with so-called global ratings of language production, in which raters may tend to develop their own internal componential rating criteria, which differ from rater to rater and within a single rater on different occasions" (p.220-221).

In relation to this problem, Fulcher (2014) states that while the scales may be inadequately specified, it may be meaningful for experienced raters who have been trained and socialized in the use of the scale. 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. Although rater training can reduce random error in rating, it does not remove completely the differences in severity between raters. So, the question –to what extent the scale was meaningful when it was separated from the training needs to be further investigated. Taylor (2011) with reference to Myford and Wolfe (2003, 2004) provides a useful summary of rater effects where the training was exempted:

(1) "The 'halo' effect, which occurs when the assessment of one trait (e.g. grammar and vocabulary) influences the evaluation of a different, conceptually independent trait (e.g. discourse management)" (p.209).

(2) “The ‘central-tendency’ effect, which refers to the overuse of the middle category of a rating scale and the avoidance the extremes” (p.209).

(3) “The ‘randomness’ effect, where raters show a tendency to apply one or more categories of the rating scale in an inconsistent way, showing random variation” (p.209).

Even very experienced raters can be affected by these problems. The investigation of each effect was beyond the scope of this study. The use of sophisticated statistical methods, such as multi-faceted Rasch techniques, can better allow us to understand this aspect of rater behavior. Even though rater training and standardization are beneficial in reducing rater variability, they still may “mask problems with the wording of bands in the scale by creating the illusion of psychological reality through high rater reliability” (Fulcher, 2014; p. 97). At this point, it should be stated that the rating scale used in the present study was a-theoretical in nature. That is, it was not based on “empirically validated descriptions of language proficiency”. It was, rather, an attempt to present a working framework that was needed within a computer-based testing context. In the absence of a theoretical model on which to base the framework, the study adopted the scale developed in the institution. The most probable reason for achieving a good degree of inter-rater reliability can be the familiarity of the scale by the raters in study. They were probably capable of reliably using such a rating scale due to their understanding of the criteria.

The type of the rating scale preferred in the study might also have an effect in the results of inter-rater reliability. The goal of the rating scale is to guide the rating process (Fulcher, 2014). There are different types of rating scales used for the assessment of speaking (e.g. holistic, analytic). The rating scale used in the study was analytical one. Although research addresses holistic scales to be used by well-trained raters, this study adopted an analytic scale for the assessment. As Fulcher (2014) cautions that a single score as in the holistic scales may not do justice to the complexity when speaking is concerned. Having multiple observations, on the other hand, can lead to score reliability be enhanced (Taylor & Galaczi, 2011). So, the analytic scale as a type of rating scale was preferred since “more scores, or pieces of information about a test taker’s ability will translate to greater consistency” (Carr, 2011, p.133). Another point in relation to the use of the rating scale in the present study was that separate

judgments were required for each task in the test. The reason was that scoring each task separately would lead to greater reliability and dependability since it provides greater control over rater behavior (Carr, 2011).

Having two raters instead of one might also be one of the reasons in achieving a good reliability. From the studies of inter-rater consistency, Mullen (1980) indicates that two raters are needed in the assessment of speaking since the raters have their own way of patterns for rating. As argued in Fulcher (2014), the use of double rating can avoid the potential affect 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).

### 5.3. Order-effect

In order to minimize a possible test order effect, test takers in the present study were assigned to two groups, each group taking the test in a different order. The results of the ANCOVA statistics showed that test order effect was present. On the whole, the group which took the COPTEFL first were found to do better on the face-to-face speaking test, and the group which took the face-to-face speaking test first did better on the COPTEFL. Obviously, 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 a 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 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 lies 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. As an implication of this finding, Zhou (2009) suggested that rather than adopting a counterbalanced design, further studies may conduct the face-to-face mode first, followed by the computer mode and "as such, the practice effect that has led to the decrease of data used in the statistical analyses in this study may be avoided" (p.116).

### 5.4. Score Comparability

After the investigation of the test order effect, correlation coefficient and paired-sample *t*-test statistics were computed for the score comparability analyses between the two test modes. The findings of the paired-sample *t*-test by test delivery mode did not reveal a statistically significant difference between the mode means for either total scores or each task type. 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 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 (Thompson, Cox & Knapp, 2016; Zhou, 2015a; Mousavi, 2007). 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). Event 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. The correlation coefficient score in the study was lower than Joo (2008) and the same with Mousavi (2007) and slightly higher than Öztekin (2011) and Jeong (2003).

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 relationship between the two delivery modes and high correlations might be accomplished even if 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).

Another possible interpretation of the lack of significant differences in the mean scores across modes 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. He found in his study that the test takers performances were similar across the tasks based on the scoring criteria employed in the test. 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 taker 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 in 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. This problem might be because

of the heavy use of adverbs or quantifiers in the scale (Carr, 2011). The little meaning difference between quantifiers may lead to ambiguity in descriptors which therefore results in raters' adjusting variability in performance and applying their own understanding consistently in both tests. In this case, it seems crucial to have double-ratings to avoid each rater's generating unreliable results. From another perspective, according to Kohen (1999, 2000) cited in Mousavi (2007), there are four aspects of differences between two modes; "(1) differences in test questions, (2) differences in test scoring, (3) differences in test conditions and (4) differences in examinee groups" (p.179). In the present study, the tests differed only in delivery mode. The same question bank was used for tasks and rating scales as well as the raters and the test taker groups were the same for each test mode. In relation to this, as stated in Zhou (2015a), "if delivery mode is considered one dimension of the speaking task, the unobserved differences in test scores is not surprising" (p.12). Despite any unobserved differences in the ratings between the two tests, Zhou (2015b) remarks that it is also possible for test takers to perform differently with regard to the certain categories of assessment. Ortega (2005), for example, found that when a listener was present, some test takers produced simpler language output, made fewer self-correction, and stopped less to prevent confusion. However, in the absence of a listener, they used more complex and accurate but less frequent language. Similarly, the test takers in the present study may have performed differently across two modes of tests. In other words, the language output elicited via the COPTEFL might be different from that elicited through the face-to-face speaking test and the two tests could be testing distinct components of oral proficiency.

Even though numerous research revealed high correlations between test scores, according to what O'Loughin (1997) argues, getting similar scores on two modes may not suggest that the two test modes measure the same kind of ability or that the ability is measured equally, which indeed means that one of the modes might be lacking construct validity. In general, test scores are considered to be crude indications of oral proficiency, since they are depended on raters' impression alone, aided by a rating scale (Zhou, 2009; Elder & Iwashita, 2005). However, discourse analytic measures have been regarded to be relatively more objective indicators of actual performance (Ellis & Barkhuizen, 2005). From the various discourse analyses of the comparisons of the two test modes, Shohamy (1994) reveals that there were differences between the

communicative strategies and discourse features produced in direct (OPI) and semi-direct (SOPI) testing of oral proficiency. She discusses that during the OPI, it was possible for test takers to note the reactions of interviewers and to adapt their speech accordingly by a variety of strategies. She further claims that the presence of an interviewer on the OPI was very likely the cause of language output that was more conversational and intimate, while language output on the SOPI had "tape-like" discourse involving a limited number of speech functions and consequently, the language produced on the OPI turned out to be a "conversational interview" and on the SOPI a "reporting monologue". Shohamy's (1994) explanation of the various differences between the two modes is that the test mode or test context (i.e. face-to-face mode or tape-mediated) rather than the elicitation tasks, is the most determinant of the type of language produced: "The different oral discourses that are produced by the two tests are a result of the different contexts in which the language is elicited. Context alone seems to be more powerful than the elicitation tasks themselves" (p.118).

Similarly, O'Loughin (1997, 2001) also asserts that the spoken interaction of two people is jointly constructed so it is basically different from communicating with a machine. Therefore, researchers (O'Loughin, 1997; 2001; Shohamy, 1994; Clark, 1979) caution against using direct and semi-direct oral proficiency tests interchangeably. Obviously, as Öztekin (2011) puts forward there might be plausible reasons to claim semi-direct speaking tests to be a form of testing that lacks reliability and concurrent validity since there is a possibility that results from them may not match with those from their face-to-face counterparts with regard to content, even if the test scores correlate. Yet, results from Quaid (2018) have shown that the reverse may be true for the general implication in Shohamy's (1994) study. According to what Quaid (2018) claims, task effect is highly significant variable which needs to be controlled for in comparative studies. In his study, Quaid (2018) investigated test taker spoken output equivalence in an identical direct and semi-direct speaking test. The results indicated that the test taker performances were highly comparable between modes. Evidence from discourse analyses of the comparisons in their study showed that Shohamy's (1994) argument that "the oral discourses that are produced by the two tests are a result of the different contexts in which the language is elicited" might not be true and according to him, Shohamy's (1994) findings might be resulted from task and interlocutor effect. Therefore, he concluded that use of identical task type and content should be given

fuller consideration. Similarly, in his study Fulcher (2003) concluded that research that have revealed significant differences in performance across tasks used maximally different tasks. There is supporting previous related research to this view (Choi, 2014; Zhou, 2008) in that when the interlocutor effect is controlled and closely matched monologic test tasks are used, similar underlying factor structures are found between the two modes. Also the findings in relation to the discourse analysis are not congruent with the prediction that monologic tasks delivered via semi-tests assess monologic ability whereas those in the face-to-face mode tap interactive ability (Zhou, 2009). This prediction has been based on the findings of O'Loughlin (2001) that lexical density revealed on monologic tasks between the two modes were significantly different. Contrary to the suggestion provided by O'Loughlin (2001), as Zhou (2009) ascertains, the simpler feedback provided by the interviewer does not seem to result in a certain degree of interaction. The focus in a face-to-face speaking test is on the successful elicitation of the language rather than on successful conversation. Therefore, an interview is distinguished from a conversation by 'asymmetrical contingency': "The interviewer has a plan and conducts and controls the interview largely according to that plan" (van Lier, 1989, p. 496). However, a conversation is characterized by "face-to-face interaction, unplannedness (locally assembled), unpredictability of sequence and outcome, potentially equal distribution of rights and duties in talk, and manifestation of features of reactive and mutual contingency" (van Lier, 1989, p. 495).

On the whole, as Amengual-Pizarro and Garcia-Laborda (2017) report in relation to the evaluation of the speech samples, the findings of previous research are encouraging in that the semi-direct testing can be used as an efficient complement to face-to-face assessment.

Several other researchers have contrasted the relative advantages and disadvantages of direct and semi-direct tests of oral proficiency. Generally in these studies, it is normally the validity of direct measures of oral proficiency which is taken for granted and that of semi-direct versions which need to be established. Accordingly, semi-direct tests lack sufficient predictive validity since they do not reflect the way most people would communicate in a real context whereas direct tests allow the test taker to talk to a live person, just as they communicate in real life. Talking to a machine has not only lower face validity but also lower construct validity than talking to a real person, due to the high degree of artificiality brought to the test in the former (Qian,

2009). However, the goal of mirroring real life in a test has been criticized as naive since no test setting exactly resembles its real-life counterpart (Bachman, 1990). The study by Hoejke and Linnell (1994) also highlights the fact that authentic language use is not necessarily guaranteed in oral proficiency testing, irrespective of whether a direct or semi-direct mode is preferred. In their study, they showed that the interaction in the direct test (OPI) still differs considerably from that of the classroom context since it occurs between two speakers and it is the interviewer who controls the discussion. Clark (1979) also acknowledges the inability of the face-to-face speaking tests to accurately reflect 'real life' conversational settings: "In the interview situation, the examinee is certainly aware that he or she is talking to a language assessor and not to a waiter, taxi driver, or personal friend" (p.38). However, once the need for an oral proficiency test has been established, which of the two tests should be preferred? In relation to the issue of choosing from either direct or semi-direct procedures, Shohamy (1994) reports that testers need several sources from which they can choose the most relevant one for the specific situation. The purpose and intended use of tests and their results are addressed as a variety of contextual variables that the answer of this question is depended on. According to Clark (1979), it should be the one "best meets both the informational needs of the test user and the practical limitations inherent in the testing situation" (p.37). In a similar vein, Galaczi (2010) reminds us the "fitness for purpose" in establishing construct validity with reference to the decisions made on the basis of test scores: "Tests are not just valid, they are valid for a specific purpose, and as such, different test formats have different applicability for different contexts, age groups, proficiency levels, and score-user requirements" (p.26). According to Kenyon and Malabonga (2001), if the assessment of interactional competence is critical, it may be a better option for a while to provide face-to-face speaking tests before it can be replicated with technology. However, in instances where the criteria adopted to evaluate test takers' performances do not differentiate on their interactional abilities, it may not be necessary to use the labor-intensive face-to-face speaking tests, irrespective of test takers' feelings about the nature of the assessment. Also, if it is not practical or feasible to offer a one-on-one interview, it may not be again necessary to administer a labor-intensive face-to-face speaking test. For some cases, they say that it may be imperative that test takers be conducted a common assessment and that differences between the skills of different interviewers to administer a face-to-face speaking test cannot be

tolerated. In such cases, they report that technology-based assessments may be the best option. This view is also expressed by Stansfield, Kenyon, Paiva, Doyle, Ulsh and Cowles (1990) in that under certain circumstances semi-direct versions (i.e. PTS; Portuguese Speaking Test) may be preferable even when an interviewer is available. One such situation may be when the quality of a rater is a major concern, for example when an important decision will be made based on the oral proficiency score awarded to a test taker, it may be preferable to use the semi-direct version. This is because the semi-direct versions are carefully constructed and trialed with a concern that questions are of equivalent difficulty. However, face-to-face speaking tests are always different in content, even when expert interviewers conduct the test. At this point, as Amengual-Pizarro and Garcia-Laborda (2017) indicate computer-based speaking tests may provide a promising procedure to assist face-to-face speaking tests, contributing to decision-making in multiple educational and occupational settings. As for Shohamy (1994), a valid speaking testing procedure may require the use of both types of tests depending on the purpose of the test and the intended use of its results. This view is also expressed by Luoma (1994) in that the best solution could be to incorporate both test formats given their respective strengths and weaknesses. Therefore, it is believed in this view that both tests should be perceived as complementary rather than as competing alternatives. From another perspective, Stansfield and Kenyon (1992) assert that test delivery modes may be better suited for particular task types and particular proficiency levels. Accordingly, the SOPI may be a better format to elicit and assess extended speech and longer turns rather that short responses that are more typical interaction between two people. In this regard, the SOPI format may be

> "somewhat more friendly for the examinee at the Advanced level or above on the ACTFL scale than for the Intermediate level examinee, because Advanced level speakers are able to use paragraph length discourse. It may also be the case that the kind of interaction provided by the face-to-face interview is most appropriate for the Novice and Intermediate level examinee" (p.363).

However, as Stansfield and Kenyon (1992) state these are merely hypotheses that need to be further investigated through research. It is important to keep in mind that some of the studies cited above were tape-based semi-direct tests and consequently, results of these studies may not directly relate to the current study.

Thus, in making decisions as to whether to use one test or another, consideration must be given not only to concurrent validation, using correlations, but also the validity of the test from multiple perspectives. The current study suggested that test taker

attitudes might represent an important source to obtain deeper understanding to the construct 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 validity. Previous research findings also support this view in that

> "If test takers have negative attitudes to the test then they are less likely to perform to best of their abilities. This has obvious implications for test validity. If test attitudes interfere with test performance this may result in unwarranted inferences being drawn from test scores" (Elder, Iwashita & McNamara, 2002, p.350).

Therefore, it seems obvious that what tests measure should be investigated from multiple perspectives in which test taker attitude towards test conditions is also a concern. Focusing on test takers' –who are one of the main users of the tests- attitudes provides further insights to decision makers in selecting the most appropriate test for their needs in given settings and given purposes. In the present study, therefore, test takers' attitude towards test conditions was also addressed. In order to explore it, an open-ended questionnaire was administered to the test takers to (1) record their reactions in relation to the COPTEFL system and (2) compare both test delivery modes.

### 5.5. Test Taker Attitudes

The results showed that test takers' preferences in relation to the test delivery modes varied widely. 20 (%36) of the comments were related to the positive sides of taking a face-to-face speaking test. Students who favored it referred to the advantages of the presence of an interviewer, the nature of the talk and the use of time. 35 (%64) of the comments, on the other hand, were related to the positive sides of taking the COPTEFL. Students who favored the COPTEFL referred to the advantages of the absence of an interviewer, better control of the pace of test taking and its being a more fair test. Qualitative analysis of the data revealed that test takers had favorable attitudes towards the COPTEFL in many aspects and 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 COPTEFL 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 fort" (p.123).

As part of the endeavor in providing evidence for the test taker attitudes towards test delivery modes, Research Question 3 investigated the reactions of test takers in relation to the COPTEFL. Test takers' responses to the open-ended question *"How do you evaluate the COPTEFL system as a speaking test delivery medium?"* yielded deeper insights into their attitudes with regard to the newly developed test instrument, the COPTEFL. In general, test takers reacted positively to the features of the COPTEFL. Specifically, they reported on the simplicity and clarity of its interface design. During the design stage of the COPTEFL system, potential distraction features i.e. unnecessary buttons, unconventional color combinations or poor quality audio or video were taken into account to minimize the effect of such construct-irrelevant factors so that test takers' performances would not be affected adversely. As Mousavi (2007) states such construct-irrelevant factors in the interface design may inadvertently influence test taker performance in the CBT and invalidate the results. The qualitative data analysis of the results revealed no undue interference of such factors or confusion in the interface design at least. This finding highlighted Fulcher's (2003) view that "interface development and design is extremely important in computer-based testing, where usability problems may constitute a threat to construct validity" (p. 384). Another factor which was found to contribute to the test takers' positive attitudes towards the COPTEFL was the ease and the efficiency in its operation. Of particular, test takers commented on the ease and the efficiency of running the instructions, tasks and the speed of delivery. A few of them, on the other hand, reported that microphone quality, the countdown timer on the screen and the transition between instructions and the tasks should be re-considered. All in all, it was understood from the responses that test takers favored the COPTEFL and found it reasonable to handle. According to Amengual-Pizarro and Garcia-Laborda (2017), this finding is important since computer familiarity

and the other features of the computerized context (e.g. computer anxiety) may have an influence on test takers' performance and prevent them from demonstrating their real oral proficiency.

As part of the Research Question 3, test takers were also asked to compare both test delivery modes and express their opinions on the form of each assessment. Similar to previous research findings (Amengual-Pizarro & Garcia-Labora, 2017; Öztekin, 2011; Qian, 2009), the face-to-face speaking test was found to be more efficient to evaluate their oral skills. Among the prominent reasons cited for this finding is the "unnatural" structure of the CBTs as opposed to the interactive face-to-face speaking tests bearing a communicative nature. Large of these research has presented results in support of the view that test takers mainly regard oral communication as human phenomenon. The main reasons provided by test takers in this regard were mostly related to the interactional nature of conversation (i.e. use of real or "authentic" language, importance of non-verbal communication, and having feedback), which could not be effectively elicited by technologically mediated assessment (Amengual-Pizarro & Garcia-Labora, 2017). In the findings of the present study, the face-to-face speaking test drew the most positive results on this aspect and test takers highlighted its intrinsically social and interactional nature compared to the COPTEFL. The main reason why the majority of the test takers viewed the face-to-face speaking test as a more effective evaluation of their oral proficiency was attributed to the presence of an interviewer. That is, they felt more natural talking to an interviewer by using non-verbal communication i.e. gestures and facial expressions and feedback. They appreciated the possibility of asking to repeat and clarify the questions they found difficult to understand and having some feedback from the interviewer to change or correct their wrong responses. Further, they stated that an interviewer could encourage them to get to talk in case they had trouble in answering the questions and this made them feel at ease and relaxed. However, the presence of an interviewer can also negatively affect test takers' performance and add further pressure (Amengual-Pizarro & Garcia-Labora, 2017). Test takers may feel more relaxed when there is no one listening to them, as in the CBT forms of assessment. In line with this, the most frequently cited positive comment about the COPTEFL was that it made them less anxious. The findings revealed that those who had more favorable attitudes towards the COPTEFL cited the ideas favoring the face-to-face test as the negative influences to show their willingness

to have a computer-based speaking test. This finding was in line with what Zhou (2009) also indicated. According to the analyses of the findings in Zhou (2009), it was revealed that the presence of the interviewer was the main reason for the participants' higher level of anxiety in the face-to-face speaking test. It was found that the test takers who felt more anxious in the computer-based speaking test also reported the presence of the interviewer, but in that case, they focused on interviewers' helpful role in making them feel at ease. In the current study, those test takers who especially favored the COPTEFL, mostly said that they felt less anxious when they were talking to a computer in the COPTEFL rather than an interviewer in the face-to-face speaking test. According to them, talking to a real person increased their level of anxiety and made them feel that they were being judged for their mistakes. They felt that the computer's nonjudgmental presence allowed them to be more relaxed and better able to demonstrate their speaking ability. These results are consistent with those in Thompson, Cox & Knapp (2016), in which comments from test takers showed that the presence of the interviewer was a reason for their higher level of anxiety. According to them, they felt less nervous with the computer mode. As noted by Zhou (2009), these results suggested that when interpreting the test takers' attitudes towards testing speaking, their personal characteristics i.e. degree of anxiety need to be taken into consideration.

One could think that the differences between test takers' attitudes towards the test delivery modes might have been affected by their test scores. But test takers in the present study were awarded scores after they expressed their attitudes about the tests and therefore, this made it impossible for test scores to have an impact on test takers' attitudes.

Regarding the computer mode, it is worth noting that several test takers mentioned being able to see preparation and response time as a factor for the affective appeal of the computer mode. This is consistent with a proposition made in Kenyon and Malabonga (2001), that is, providing test takers with more control over an assessment has a positive influence on their affect. This is also consistent with the findings in Zhou (2009) that several test takers in that study reported being able to control preparation and response time as a positive side of the computer-based speaking test. However, in the present study, specifically the response time, which was shown with a countdown timer on the screen, may have led some test takers to feel nervous because of time pressure and prevented them from demonstrating their best performances. The findings

showed that some of the test takers found it as stressful to see a clock which reminded them how long they had left to speak for the given task. Also, a few of the test takers reported that the time given for the preparation and response was not enough to answer the questions. This finding suggested that variations of longer or shorter time periods must be investigated with different groups of test takers in order to find optimal time required for both preparation and response periods.

Apart from its providing better control of the pace of test taking, the test takers also referred that the COPTEFL was a more fair test compared to the face-to-face speaking test. They claimed that in general, their test scores vary according to who conducts the test. In their opinion, if a difficult interviewer administered the test, they would have difficulty in performing well. But, if an easy and familiar interviewer asked the questions, they would get a higher score because the interviewer supported enough. However, in such a computer-based speaking test, they indicated that there would be no worry about who the interviewer would be and whether they could get provided scaffolding or not. In relation to the test fairness, one of the test takers also reported that the use of the COPTEFL could help prevent raters from being affected by their personal characteristics, which was often believed to be the case in human-scored tests (Amengual-Pizarro & Garcia-Labora, 2017). Therefore, there were perceived advantages of the COPTEFL in terms of objectivity of the test scores. This opinion has given support to the argument by Stansfield and Kenyon (1992) in that the semi-direct oral proficiency test, in which there is no interviewer, is 'fairer' than face-to-face speaking test. The variability in interviewers' administration of the test observed in many studies led the researchers to take into account the need for interviewer training. Such training is considered as the way to ensure greater uniformity in interviewers' behavior and, therefore, more consistent ratings (Szmerdt-Chandler, 2011). Training interviewers on how they should and should not behave was crucial for the purposes of the study. In order to compare the scores obtained by two modes, it was necessary to arrange the same administration standards for both tests. Since the COPTEFL was lack of an interviewer, the face-to-face speaking test needed to be controlled by the influence of the interviewer on test scores. In order to avoid the threat to fairness to all test takers, an interviewer-training session was held. In this meeting, what types of interviewer support would be allowed was told. They were required not to provide scaffolding in order to ensure the comparability of test modes. A highly-structured script was given to

each interviewer so that each would administer the test in the same way. The purpose of the training therefore was to help lead them to apply the test in the same manner since the effect of the interviewer on the test administration may lead variation in the way tasks are provided and any variation may impact test takers' performances (O'Loughlin, 2001). As Lazaraton (1996) states:

> "In fact, the achievement of consistent ratings is highly dependent on the achievement of consistent examiner conduct during the procedure, since we cannot ensure that all candidates are given the same number and kinds of opportunities to display their abilities unless oral examiners conduct themselves in similar, prescribed ways" (p.19).

All in all, the findings of the study in respect to test taker attitudes implied that test takers' personal characteristics and preferred interpersonal style must be taken into account when selecting an assessment format. According to what previous research suggests, test takers' personal characteristics such as level of extroversion or introversion can influence their performances in oral assessments (O'Sullivan, 2000). Mousavi (2007) also cautions that reactions to the test conditions might also be the results of individual differences between test takers and consequently, these reactions may take various forms at different levels of proficiency and with test takers from different language and cultural backgrounds. These discrepancies, however, were not reflected on the responses made by the test takers in the present study. Because test takers' preferences varied widely, preparing them to take oral proficiency tests is crucial. The participants in this study did not receive any prior information in relation to the COPTEFL format. Perhaps, if they were familiar with taking a test in computer mode, their opinions would change. The greater use of the COPTEFL would lead to the more positive attitudes towards its employment in proficiency tests. However, this assumption requires a further consideration and research on how test taker characteristics and behavior interact with the COPTEFL.

To sum up, the major advantages of the face-to-face speaking test were found related to the presence of an interviewer which enhances interaction and encourage them to feel at ease and talk more. They appreciated the possibility of some warm feedback on the part of the interviewer, which was lacking in the COPTEFL. On the other hand, the presence of an interviewer was also addressed as a negative aspect of testing speaking that increasing the test takers' anxiety level. According to the test takers who favored the absence of an interviewer, the COPTEFL produced more

reliable results since their performances were not affected by the interviewers' inferences during the test.

This chapter summarized the major findings of the present study and discussed the relevant issues in respect to each research question. The following chapter presented the conclusions, implications and limitations of the study, and suggestions for further research.

# 6. CONCLUSION

## 6.1. Summary of the Study

In order to test oral proficiency appropriately and efficiently, various language tests have been developed over the years. For example, the American Council on the Teaching of Foreign Languages (ACTFL) developed a direct face-to-face interview format (the Oral Proficiency Interview, OPI) to evaluate learners' second language competence. The OPI is conducted by trained interviewers and raters, and widely available to test takers in need of rating their competency. Concerns, however, have been raised about potential confounding effects of human testers and demanding logistics requirements for administration of the test (Yu, 2006). In order to offer a solution for the practical constraints and eliminate the need to sustain a costly and labor-intensive direct format, a semi-direct testing format (the Simulated Oral Proficiency Interview, SOPI) (using an audio-taped recorder and printed stimuli and recording test takers' responses) was subsequently employed to deliver the tests. The SOPI carefully follows the OPI format in that it begins with a warm-up and then the difficulty of tasks increases during the assessment. This approach eliminated the possible confounding effects of human tester and logistics requirements for test administration (Stansfield, et. al, 1990). Since then, there have been several efforts to develop semi-direct tests of oral proficiency in foreign languages. The need for better test delivery formats drew attention to advanced computer technology and a computer-based alternative, the Oral Proficiency Interview-Computer (OPIc) was developed. Similar to the OPI, the OPIc produces a ratable speech sample that is subsequently double-blind-rated by at least two certified OPIc raters (Thompson, Cox & Knapp, 2016). The OPIc has often been expressed as "[A] different modality of the ACTFL OPI and not a different assessment" (Surface, Poncheri & Bhavsar, 2008, p.8). Although many research findings supported the comparability of the OPI and the OPIc (i.e. Surface, Poncheri & Bhavsar, 2008), the findings suggested some areas of obvious concern, such as test-retest reliability, inter-rater reliability, and rater agreement (Thompson, Cox & Knapp, 2016).

As demand for oral language proficiency tests continue to grow, the integration of computer technology into the speaking tests in foreign languages is gradually becoming more involved in test administrations (Amengual-Pizarro & Garcia-Laborda, 2017). Consequently, oral proficiency tests have been delivered in a various computer-

based test formats (i.e. OPIc, TOEFL IBT). However, numerous concerns have been addressed over the validity of such tests (Zhou, 2015a). Computer-based speaking tests are thought to restrict the range of task types to elicit language samples and narrow down the test construct because of the lack of an interaction between individuals and therefore, they are reported to be not able to capture the complexity of natural language use (Amengual-Pizarro & Garcia-Laborda, 2017). However, there are also research findings revealing that despite the difficulty of capturing human oral language interaction (Amengual-Pizarro & Garcia-Laborda, 2017; Douglas & Hegelheimer, 2007; Kenyon & Malabonga, 2001) computer-based tests can provide a valid measure of oral proficiency in English and be a suitable testing method for the assessment of oral skills. Specifically, there are some authors who strongly support the use of computer technology in educational settings (Garcia-Laborda, 2007, Qian 2009). According to them, computer language testing can be served as a feasible alternative to other traditional methods of testing oral skills (i.e. face-to-face speaking tests) because it facilitates test administration and delivery by reducing testing time and costs. Over the few years, the number of institutions that use computers as a method for testing oral proficiency has increased dramatically (Garcia-Laborda, Magal Royo & Barcena Madera, 2014). However, this situation is not the same everywhere. While countries like the United States or the United Kingdom are spearheads of this tendency, others like Turkey, lag way behind. In order to add to the literature in the context of the study, in the present research, it was aimed at developing and validating the Computerized Oral Proficiency Test of English as a Foreign Language, COPTEFL. The purposes of this study were then, firstly to develop the COPTEFL and then to shed light on the validity and usefulness of a newly developed the COPTEFL. For the purpose of the development and validation, the stages of constructing the COPTEFL were explained in a detailed way and the scores obtained from the COPTEFL were compared with those of the face-to-face speaking test. Finally, as for the usability analysis, test takers' attitudes towards the COPTEFL were explored. With these aims, the study addressed the following research questions: (1) What are the cognitive demands of the speaking test tasks? (2) What are the contextual characteristics of the speaking test tasks? (3) 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? (4) What are the attitudes of test takers in relation to the test delivery modes?

In order to investigate the research questions, the study employed a "mixed methods explanatory sequential design" (Creswell, 2012). In this design, the quantitative data collection and analysis were followed up with the qualitative data collection and analysis. The results from the study depicted to the extent to which the COPTEFL can serve as a viable substitute for the conventional face-to-face speaking test. For the purposes of this research, a software program was designed, developed and administered along with a face-to-face speaking test. The study tested the performance of a group of test takers on two delivery modes. It was aimed to find out whether scores obtained from the administration of the COPTEFL was comparable to those obtained from its conventional counterpart. That is, this research was correlational in nature.

The study was employed at Anadolu University School of Foreign Languages (AUSFL), Turkey, in which the aim is to equip the learners with necessary language skills to follow the academic education in their departments. 45 students participated in the study at the time of data collection and participation was on a voluntary basis. There were 8 item writers who were also the raters. They were professional instructors who work full-time for testing institution at AUSFL. In the editing committee, there were four experts who were asked for opinions about the written items. The data for this study were gathered from multiple sources including (1) the COPTEFL scores, (2) the face-to-face speaking test scores, and (3) the responses to open-ended questions in relation to the test taker attitudes towards test delivery modes. Before the main data collection procedure, test tasks and the COPTEFL system were trialled and evaluated through pilot tests. The final version of the task types and the COPTEFL system were based on these pilot tests. For the main data collection, a counterbalanced design was employed. In this design, test-takers were assigned at random to two testing modes; (a) the COPTEFL first and the face-to-face speaking test second or (b) the face-to-face speaking test first and the COPTEFL second. With this design, it was aimed to find out whether the test in one mode followed by the other affected the score for the second mode. To measure the effect of delivery mode and the mode-by-order interaction statistically, analysis of covariance (ANCOVA) with within-subject effect was run on total scores. After that, in order to assess the consistency between judges' ratings, inter-rater reliabilities were calculated for each test delivery mode. To examine the inter-rater reliability, two-way random absolute agreement intra-class correlation coefficient (ICC) was run. For the investigation of the equivalence of test scores across modes, the

magnitude of raw scores was compared by means of ICC. For better interpretations of the equivalence of test scores across modes, 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 face-to-face speaking test. Finally, as for the test taker attitudes, the responses to open-ended questions were analysed based on the qualitative content analysis scheme of Creswell (2012).

The findings of the study revealed that there was a significant interaction of test order and test scores. The results suggested that both groups did better in their second test than in their first test. This showed that there was a practice effect in general and the test takers improved their speaking practice within the period between the administrations of the two tests. ICC results for inter-rater reliabilities across test delivery modes showed that the ICC score for the face-to-face speaking test (ICC= .889) was slightly higher than the score in the COPTEFL (ICC= .806). These ICC values between .75 and .9 indicated a good reliability for both tests. ICC results of inter-rater reliabilities for each task type across test delivery modes indicated a significant direct relationship between inter-rater reliability scores for each task type across test delivery modes that those who got higher scores from a task in the COPTEFL by the rater also got higher scores from the same task in the face-to-face speaking test by the inter-rater, or vice versa. The correlational coefficient between the scores by two variables as the COPTEFL and the face-to-face speaking test was found to be .63, which is considered a positive moderate index of reliability. This result indicated a direct relationship between the scores across test delivery modes that those who got higher scores from the COPTEFL also got higher scores from the face-to-face speaking test, or vice versa. As for the scores from each task type, a low degree of ICC was found between COPTEFL opinion task scores and face-to-face speaking test opinion task scores. For the comparison task scores, there was a significant moderate degree of ICC between the COPTEFL and the face-to-face speaking test. 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. Furthermore, mean score differences showed that there was not a statistically significant difference between the COPTEFL and the face-to-face speaking total test scores. When the mean scores of each task type were investigated, the findings showed that there was not a statistically significant

difference between the COPTEFL scores and the face-to-face speaking test scores for task types. In order to shed lights on the preferences of test takers with regards to delivery modes, an open-ended questionnaire was administered to the test takers to (1) record their reactions in relation to the COPTEFL system and (2) compare both test delivery modes. In general, test takers reacted positively to the features of the COPTEFL. Specifically, they reported on the simplicity and clarity of its interface design. They also commented on the ease and the efficiency of running the instructions, tasks and the speed of delivery. A few of them, however, reported their concern on microphone quality, the countdown timer on the screen and the transition between instructions and the tasks and stated that they should be re-considered. All in all, it was understood from the responses that test takers favored the COPTEFL and found it reasonable to handle. Test takers were also asked to compare the advantages and disadvantages of taking a computer-based test with taking a conventional face-to-face speaking test. The findings revealed that those who favored the face-to-face speaking test referred to the advantages of the presence of an interviewer, the nature of the talk and the use of time. In the majority of cases, these students reported that taking a face-to-face speaking test feels more like a real speaking test in which they interact with a real person as in the daily life conversation. However, those students who favored the COPTEFL cited the same reason as their preference for a computer-based speaking test. According to them, talking to a real person increased their level of nervousness and made them concentrate on their mistakes since they feel like being judged by the interviewer. In majority of cases, these students referred to the advantages of the absence of an interviewer, better control of the pace of test taking and its being a more fair test. These findings showed that students' personal characteristics and preferred interpersonal style may be required to be taken into consideration when selecting an assessment format. In this sense, a further investigation into how personal characteristics and preferred interpersonal style and attitudes towards test delivery modes is needed to find out how they are interacted with each other.

### 6.2. 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 of 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 a 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.

Another limitation of the study is the monologic nature of the tasks in the speaking tests. Due to the limitation of the computer-based speaking tests, interactive tasks that require test takers to interact with interviewers were not included in the study. The attempt was to equate the two testing formats in order to get solid and valid data about test takers' performances on both tests by using the similar tasks and the same elicitation techniques. Thus, this research only focused on monologic tasks, which elicit performance representing a narrow interpretation of oral proficiency. Therefore, caution is advised in generalizing the findings to interactive tasks.

Lastly, time constraints were also among the limitations of the present study. Designing test specifications, test tasks and developing the COPEFL system and piloting of the test tasks and the instruments and finally, administering the both tests were completed in a short period of time. Within this time period, the COPTEFL should have been provided to the students for a few trials to get used to it before the actual test.

### 6.3. Implications and Suggestions for Further Research

Several implications can be drawn from the results of the study in relation to language testing and test development processes. The COPTEFL has been presented as a robust tool that would enable the test administrators and language testers to deliver the English language oral proficiency tests via the Internet. It is a newly developed testing format, which is still in its infancy and requires further trials with larger groups of participants and consequently, a number of years to become fully developed. However, the administration of the COPTEFL showed that implementing a computer-based speaking test as part of the oral proficiency testing is feasible and promising for promoting language testing conditions and overcoming the practical difficulties

especially regarding personal variables of the interviewers and raters involved in face-to-face speaking tests. The results of the experiment 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 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 new comers 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 findings of the study also provided positive evidence for the validity of the COPTEFL in terms of test scores and test usefulness. The fact that there were no statistically significant mean score differences between the COPTEFL and the face-to-face speaking test, and a positive moderate index of reliability implied that test users could use the test result from the COPTEFL to infer their scores in the face-to-face speaking test. Also, test takers appeared to be in favor of its implementation and their positive attitudes towards the COPTEFL as a testing system (i.e. the simplicity and clarity of the interface design, the ease and the efficiency in its operation) demonstrated the usefulness of the COPTEFL. Majority of the test takers agreed that they had felt no additional strain while being tested through it. The COPTEFL, in this sense, has been considered as one of the most achievable alternatives to face the challenges arising from the administration of the face-to-face speaking tests. However, from the test takers' side, the responses showed that if given the choice, most of them would prefer the face-to-face speaking test, which might be a result of unfamiliarity with the COPTEFL. This result 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. Although the COPTEFL enables tests to be done more efficiently, and practicality is a crucial aspect of test design, other factors, particularly construct validity should also be considered. Since test takers' speech samples were not compared between two modes, it could be possible that test delivery mode might be a confounding factor in modeling performance the COPTEFL and consequently, be a threat to the validity of the COPTEFL and to the interpretation of the test scores. So, this research calls for particular attention to study the psychometric qualities of computer-based and face-to-face monologic tasks and compare the underlying factor structures of these tasks between the two delivery modes. Further studies can also explore the impact of test takers' characteristics and proficiency level towards the type of test delivery mode in order to find out whether delivery modes cause the differences in test scores. Moreover, rater effects in scores can be further investigated. In this study, intra-class correlation coefficient is reported as an inter-rater reliability. However, the correlation coefficients do not take into account rater severity, that is, how severe or lenient a rater is. Further studies can use the multifaceted Rasch techniques to model the severity of raters.

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 scoring system. Finally, the limited number of research of this type in Turkey, also, may be a reason to further study in this field.

## REFERENCES


Ağçam, R. and Babanoğlu, M. P. (2016). Students' perceptions of language testing and assessment in higher education. *Üniversitepark Bülten / Universitepark Bulletin*, *5*(1-2), 66-77.

Alderson, J. C. (2000). Technology in testing: The present and the future. *System*, *28*(4), 593-603.

Alderson, J.C., Clapham, C. and Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Altmışdört, G. (2010). Yabancı dil öğretiminde nasıl bir ölçme-değerlendirmeye gerek vardır. *Kuramsal Eğitimbilim*, *3*(2), 175-200.

Amengual-Pizarro, M. and García-Laborda, J. (2017). Analysing test-takers' views on a computer-based speaking test. *Profile Issues in Teachers Professional Development*, *19*(1), 23-38.

American Council on the Teaching of Foreign Languages. (1999). *ACTFL proficiency guidelines—speaking: Revised 1999*. Retrieved December 04, 2016 from http://www.actfl.org/files/public/Guidelinesspeak.pdf.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual review of Psychology*, *37*(1), 1-16.

Aydın, B., Geridönmez, S., Polat, M. and Akay, E. (2017). Feasibility of computer assisted English proficiency tests in Turkey: A field study. *Anadolu Journal of Educational Sciences, 7*(1), 107-122.

Aydın, B., Unver, M. M., Alan, B. and Sağlam, S. (2017). Combining the old and the new: Designing a curriculum based on the taba model and The Global Scale of English. *Journal of Language and Linguistic Studies*, *13*(1), 304-320.

Aydın, B., Akay, E., Polat, M., ve 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.

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

Bachman, L. F. and Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, UK: Oxford University Press.

Bachman, L. F., Lynch, B. and Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*(2), 238-57.

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning and Technology*, *1*(1), 44-59.

Canale, M. (1983). From communicative competence to communicative language pedagogy. *Language and Communication*, *1*(1), 1-47.Canale, M., and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.

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

Carr, N. T. ( 2006). Computer-based testing: Prospects for innovative assessment. In L. Ducate and N. Arnold (Eds.), *Calling on CALL: From theory and research to new directions in foreign language teaching*. CALICO Monograph Series 5, 289– 312. San Marcos , TX : CALICO.

Cephe, P. T. and Toprak, T. E. (2014). The Common European Framework of Reference for Languages: Insights for language testing. *Journal of Language and Linguistic Studies, 10*(1), 79-88.

Ceylan, A. (2007). Ethical concerns in language testing at university level. MA Thesis. Onsekiz Mart Üniversitesi.

Chalhoub–Deville, M. and Deville, C. (1999). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics*, *19*, 273-299.

Chand, M. and Gold, M. (2002). *A programmer's guide to ADO. NET in C*. Apress.

Chapelle, CA, and Douglas, D. (2006). Assessing language through computer technology. Cambridge: Cambridge University Press.

Chapelle, C. A. (2013). Conceptions of validity. In *The Routledge handbook of language testing* (pp. 35-47). Routledge.

Chapelle, C. (2001). *Computer applications in second language acquisition*. Cambridge: Cambridge University Press.

Choi, I. (2014). The comparability of direct and semi-direct oral proficiency interviews in a foreign language context: A case study with advanced Korean learners of English. *Language Research, 50*(2), 545-568.

Clark, J. L. D. (1979). Direct vs. semi-direct tests of speaking ability. In E. J. Briere and F. B. Hinofotis (Eds.), *Concepts in language testing: some recent studies*, 35-49. Washington, DC: TESOL.

Council of Europe, (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge: Cambridge University Press.

Creswell, J. W. (2012*). Educational research: Planning, conducting, and evaluatingquantitative.* Boston: Pearson Education.

De Siqueira, J. M., Martínez-Sáez, A., Sevilla-Pavón, A. and Gimeno-Sanz, A. (2011). Developing a web-based system to create, deliver and assess language proficiency within the PAULEX Universitas Project. *Procedia-Social and Behavioral Sciences*, *15*, 662-666.

Douglas, D. (2014). *Understanding language testing*. London: Routledge.

Douglas, D. and Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics, 27*, 115–132.

Dunkel, P. (1999). Considerations in developing or using second language proficiency computer-adaptive tests. *Language Learning and Technology, 2*(2), 77-93.

Dunkel, P. (1997). Computer-adaptive testing of listening comprehension: A blueprint for CAT development. *The Language Teacher Online*, *21*(10), 1-8.

East, M. (2016). Mediating Assessment Innovation: Why Stakeholder Perspectives Matter. In *Assessing Foreign Language Students' Spoken Proficiency* (pp. 1-24). Springer, Singapore.

Elder, C. and Iwashita, N. (2005). Planning for test performance: What difference does it make? In R. Ellis (Ed.), *Planning and task performance in a second language,* 219–238. Amsterdam: John Benjamins.

Elder, C., Iwashita, N., and McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer?. *Language Testing*, *19*(4), 347-368.

Ellis, R., and Barkhuizen, G. P. (2005). *Analysing learner language*. Oxford: Oxford University Press.

Engin, A. (2012). Variation of Scores in Language Achievement Tests according to Gender, Item Format and Skill Areas. MA Thesis. Bilkent Üniversitesi.

English Speaking Proficiency Testing Academy (2002-2005). On-line resource. Available at http://www.espt.org

Farhady, H., Jafarpur, A. and Birjandi, P. (1994). *Testing language skills: From theory to practice*. Tehran: SAMT Publication.

Field, J. (2013). Cognitive Validity. In A. Geranpayeh and L. Taylor (Eds.), Examining listening: Research and practice in assessing second language listening, Studies in language testing, 35 (pp.77-151). Cambridge: Cambridge University Press.

Fulcher, G. (2014). *Testing second language speaking*. London: Routledge.

Fulcher, G. (2013). *Practical language testing*. Routledge.

Fulcher, G. and Davidson, F. (2007). *Language testing and assessment*. New York: Routledge.

Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing,* 20(4), 384-408.

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: the case of the First Certificate in English examination. *Language Assessment Quarterly, 5*(2), 89-119.

Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. In L. Araújo (Ed.), *Computer-based assessment of foreign language speaking skills,* 29–51. Luxemburg: European Union.

Garcia-Laborda, J. G., Magal Royo, T. M. and Barcena Madera, E. B. (2015). An Overview of The Needs of Technology in Language Testing in Spain. *Procedia-Social and Behavioral Sciences*, *186*, 87-90.

García-Laborda, J., López Santiago, M., Otero de Juan, N. and Álvarez Álvarez, A. (2014). Communicative language testing: Implications for computer-based language testing in French for specific purposes. *Online Submission*, *5*(5), 971-975.

Geranpayeh, A. (2013). Scoring validity. *Examining Listening: Research and Practice in Assessing Second Language Listening*, *35*, 242.

Göktürk Sağlam, A. L. (2016). Washback and instructional sensitivity of a theme-based high-stakes English language proficiency test. Doctorial Thesis. Yeditepe Üniversitesi.

Gömleksiz, M. N. and Özkaya, Ö. M. (2012). Yabancı Diller Yüksekokulu Öğrencilerinin İngilizce Konuşma Dersinin Etkililiğine İlişkin Görüşleri. *Electronic Turkish Studies*, *7*(2), 495-513.

Gönen, K. (2013). Teachers' perceptions of classroom-based language assessment in tertiary level English language programs in Turkey. MA Thesis. Çağ Üniversitesi.

GSE. (2016). Global Scale of English Assessment Framework. Pearson English, New York. Retrieved May 18, 2017 from http://www.english.com/gse#.

GSE. (2015). New Global Scale of English Learning Objectives. Pearson English, New York. 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.

Harlow, L. L. and Caminero, R. (1990). Oral testing of beginning language students at large universities: Is it worth the trouble?. *Foreign Language Annals*, *23*(6), 489-498.

Harmer, J. (2001). *The practice of English language teaching.* UK: Pearson Education Limited.

Heaton, J. B. (1990). *Writing English language tests: a practical guide for teachers of English as a second or foreign language*. Longman Publishing Group, Third Impression.

Hoekje, B. and Linnell, K. (1994). "Authenticity" in language testing: Evaluating spoken language tests for international teaching assistants. *TESOL Quarterly*, *28*(1), 103-126.

Hughes, A. (1992). *Testing for language teachers*. Cambridge: Cambridge University Press.

Hymes, D. (1972). *Directions in Sociolinguistics: The Ethnography of Communication.* New York: Holt, Rinehart and Winston.

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). Doctorial Thesis. University of London.

Karakuş, B. (2013). Üniversite yabancı dil hazırlık sınıflarında uygulanan modüler sistemdeki ölçme ve değerlendirme. *Eğitim ve Öğretim Araştırmaları Dergisi, 2*(4), 15-22.

Kenyon, D. M. and Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning and Technology*, *5*(2), 60-83.

Kenyon, D. M. and Tschirner, E. (2000). The rating of direct and semi-direct oral proficiency interviews: Comparing performance at lower proficiency levels. *The Modern Language Journal*, *84*(1), 85-101.

Kiddle, T. and 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., and Weir, C. J. (2009). *Examining reading* (Vol. 29). Ernst Klett Sprachen.

Koru, S. and Akesson, J. (2011). Turkey's English deficit (Policy Note). *Economic Policy Research Foundation of Turkey (TEPAV)*. Retrieved May 15, 2017 from *http://www. tepav. org. tr/upload/files/1324458212*, *1*.

Kunnan, A. J. (2013). Approaches to validation in language assessment. In *Validation in language assessment* (pp. 15-30). Routledge.

Kunnan, A. J. (1998). Validation in language assessment: *Selected papers from the 17th Language Testing Research Colloquium.*

Lazaraton, A. (1996) Interlocutor support in oral proficiency interviews. The case of CASE. *Language Testing, 13*(2),151-72.

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

Luoma, S. (2001). What does your language test measure. *Unpublished PhD dissertation, University of Jyväskyla, Jyväskyla*.

Luther, A. C. (1992). *Designing interactive multimedia*. New York: Multi-science Press Inc.

Madsen, H. S. (1983). *Techniques in Testing*. Oxford: Oxford University Press.

Malabonga, V., Kenyon, D. M. and Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing*, *22*(1), 59-92.

Malabonga, V. A. and Kenyon, D. M. (1999). Multimedia computer technology and performance-based language testing: A demonstration of the Computerized Oral Proficiency Instrument (COPI). In *Proceedings of a Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing*, 16-23. Association for Computational Linguistics.

Malone, M. E. (2007). Oral proficiency assessment: The use of technology in test development and rater training. *CALdigest, Center for Applied Linguistics:* Washington, D.C.

Messick, S. 1988. The Once and Future Issues of Validity: Assessing the Meaning and Consequences of Measurement. In Wainer, H. and H. Braun (eds.) Test Validity. 33-45.

Messick, S. (1989) Validity. In R. L Linn (Eds.), Educational measurement (pp. 13-103). New York: Macmillan/American Council on Education.

McNamara, T. (2010). The use of language tests in the service of policy: issues of validity. *Revue française de linguistique appliquée*, *15*(1), 7-23.

McNamara, T. and Roever, C. (2006). Language testing: The social dimension. *International Journal of Applied Linguistics*, *16*(2), 242-258.

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. Doctorial thesis. Griffith University.

Norris, J. M. (2001). Concerns with computerized adaptive oral proficiency assessment. *Language Learning and Technology*, *5*(2), 99-105.

Nunan, D. (1993). Task-based syllabus design: Selecting, grading and sequencing tasks. In G. Crookes and S. Gass (Eds.), *Tasks in a pedagogical context: Interacting theory and practice,* 55-68. Clevedon: Multilingual Matters.

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. (2001) *The equivalence or direct and semi-direct speaking tests.* Studies in Language Testing 13. Cambridge: Cambridge University Press.

O'Loughlin, K. (1997). The comparability of direct and semi-direct speaking tests: A case study. Unpublished doctoral thesis. University of Melbourne.

Ortega, L. (2005). What do learners plan? Learner-driven attention to form during pre-task planning. *Planning and task performance in a second language*, 77-109.

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.

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System*, *28*(3), 373-386.

Öztekin, E. (2011). A comparison of computer assisted and face-to-face speaking assessment: Performance, perceptions, anxiety, and computer attitudes. MA

Thesis. Bilkent Üniversitesi.

Purdue University. (2005). Oral English proficiency program. Retrieved May 05, 2017 from http://www.purdue.edu/OEPP/

Qian, D. (2009). Comparing direct and semi-direct modes for speaking assessment: affective effects on test takers. *Language Assessment Quarterly, 6*(2), 113–125.

Quaid, E. D. (2018). Output Register Parallelism in an Identical Direct and Semi-Direct Speaking Test: A Case Study. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, *8*(2), 75-91.

Roever, C. (2001). Web-based language testing. *Language Learning and Technology, 5*(2), 84-94.

Ruch, G. M. (1924). The Improvement of the Written Examination. Chicago: Scott, Foresman and Company.

Shaw, S. D., Shaw, D. S. and Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing* (Vol. 26). Cambridge University Press.

Shepard, L. A. 1993. Evaluating Test Validity. Review of Research in Education, 19,405-450.

Shneiderman, B. (2004). *Designing the user interface: Strategies for effective human-computer interaction: Fourth edition.* Boston: Pearson Addison-Wesley.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language testing*, *11*(2), 99-123.

Shohamy, E., Donitsa-Schmidt, S. and Waizer, R. (1993). The effect of the elicitation mode on the language samples obtained in oral tests. *Paper presented at the 15th Language Testing Research Colloquium,* Cambridge, UK.

Stansfield, C. W. (1991). A comparative analysis of simulated and direct oral proficiency interviews. In S. Anivan (Ed.), *Current developments in language testing*, 199–209. Singapore: Regional English Language Center.

Stansfield, C. W. and Kenyon, D. M. (1992) The development and validation ofa simulated oral proficiency interview. *The Modem Language Journal, 72*(2), 129-41.

Stansfield, C. W. and Kenyon, D. M. (1991). Development of the Texas Oral Proficiency Test (TOPT). Final Report.

Stansfield, C. W., Kenyon, D. M., Paiva, R., Doyle, F., Ulsh, I. and Cowles, M. A. (1990). The development and validation of the Portuguese speaking test. *Hispania*, *73*(3), 641-651.

Surface, E. A., Poncheri, R. M. and Bhavsar, K. S. (2008). Two studies investigating the reliability and validity of the English ACTFL OPIc with Korean test takers. *The ACTFL OPIc® Validation Project Technical Report*.

Szmerdt-Chandler, D. (2011). Assessing oral production. *Issues in Promoting Multilingualism*, 257-286.

Szmerdt-Chandler, Namara, T. (2000). *Language testing.* Oxford, UK: Oxford University Press.

Şahinkarakaş, S. (2012). The role of teaching experience on teachers' perceptions of language assessment. *Procedia-Social and Behavioral Sciences*, *47*, 1787-1792.

Taylor, L. (Ed.). (2011). *Examining Speaking: Research and practice in assessing second language speaking* (Vol. 30). Cambridge University Press.

Taylor, L. and Galaczi, E. (2011). Scoring validity. *Examining speaking: Research and practice in assessing second language speaking, 30,* 171-233.

Thompson, G. L., Cox, T. L. and 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. Hampshire: Palgrave McMillan.

Widdowson, H. G. (1989). Knowledge of language and ability for use. *Applied linguistics*, *10*(2), 128-137.

Wigglesworth, G. and O'Loughlin, K. (1993). An investigation into the comparability of direct and semi-direct versions of an oral interaction test in English. *Melbourne Papers in Language Testing*, *2*(1), 61-71.

Wu, W. M. and Stansfield, C. W. (2001). Towards authenticity of task in test development. *Language Testing, 18*(2), 187-206.

Yavuz Kırık, M. (2008). Yabancı dil olarak İngilizce öğretmenlerinin ölçme değerlendirme bağlamında tutum ve yaklaşımları. Doctorial Thesis. İstanbul Üniversitesi.

Yu, E. (2006). A comparative study of the effects of a computerized English proficiency test format and a conventional SPEAK test format. Unpublished doctoral dissertation, Ohio State University, Columbus.

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*, *23*(3), 489-508.

Zainal Abidin, Saidatul Akmar. (2009). Testing spoken language using computer technology: A comparative validation study of 'Live' and computer delivered test versions using Weir's framework. PhD thesis, Universiti Teknologi Mara.

Zak, D. (2001). *Programming with Microsoft Visual Basic 6.0: Enhanced edition* Boston, Massachusetts: Course Technology Thomson Learning.

Zhou, Y. (2015a). Computer-delivered or face-to-face: Effects of delivery mode on the testing of second language speaking. *Language Testing in Asia*, *5*(1), 2.

Zhou, Y. (2015b). Comparing ratings of a face-to-face and telephone-mediated speaking test. *Jacet journal*, *59*(2015), 33-52.

Zhou, Y. (2009). Effects of computer delivery mode on testing second language speaking: The case of monologic tasks. Unpublished doctoral dissertation, Tokyo University of Foreign Studies Tokyo.

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.

## APPENDIX A

## GSE Learning Objectives for Adult Learners, 2015

### GSE 51–58/B1(+): Speaking

| | |
|---|---|
| 51 | Can express opinions as regards possible solutions, giving brief reasons and explanations. ($C_A$) |
| | Can express and respond to feelings (e.g. surprise, happiness, interest, indifference). (C) |
| | Can express opinions and react to practical suggestions of where to go, what to do, etc. ($C_A$) |
| | Can make a complaint. (C) |
| | Can briefly give reasons and explanations for opinions, plans and actions. (C) |
| | Can report the opinions of others, using simple language. (P) |
| | Can express hopes for the future using a range of fixed expressions. ($CJ_A$) |
| 52 | Can repeat back what is said to confirm understanding and keep a discussion on course. ($C_A$) |
| | Can use a suitable phrase to invite others into a discussion. ($C_A$) |
| | Can speak in general terms about environmental problems. (P) |
| | Can express opinions and attitudes using a range of basic expressions and sentences. ($C_A$) |
| | Can discuss the main points of news stories about familiar topics. ($CJ_A$) |
| 53 | Can compare and contrast alternatives about what to do, where to go, etc. ($C_A$) |
| | Can use a basic repertoire of conversation strategies to maintain a discussion. ($C_A$) |
| | Can define the features of something concrete for which they can't remember the word. (C) |
| | Can develop an argument using common fixed expressions. (P) |
| | Can give a short, rehearsed talk or presentation on a familiar topic. ($C_A$) |
| | Can re-tell a familiar story using their own words. (P) |
| | Can signal that they wish to bring a conversation to an end. (P) |
| | Can ask someone to paraphrase a specific point or idea. (P) |
| 54 | Can describe basic symptoms to a doctor, but with limited precision. ($C_A$) |
| | Can relate the basic details of unpredictable occurrences (e.g. an accident). ($C_A$) |
| | Can leave phone messages containing detailed information. (P) |
| | Can make excuses using a range of polite forms. (P) |

| | |
|---|---|
| 55 | Can use synonyms to describe or gloss an unknown word. ($C_A$) |
| | Can explain the main points in an idea or problem with reasonable precision. (C) |
| | Can express their thoughts in some detail on cultural topics (e.g. music, films). ($C_A$) |
| | Can explain why something is a problem. (C) |
| | Can respond to ideas and suggestions in informal discussions. ($C_A$) |
| | Can generally follow most of what is said and repeat back details to confirm understanding. ($C_A$) |
| | Can ask for clarification of an unknown acronym or technical term used in conversation. (P) |
| 56 | Can summarise and comment on a short story or article and answer questions in detail. ($C_A$) |
| | Can give brief comments on the views of others. (C) |
| | Can summarise and give opinions on issues and stories and answer questions in detail. ($C_A$) |
| | Can give an opinion on practical problems, with support when necessary. ($C_A$) |
| | Can express and comment on ideas and suggestions in informal discussions. ($C_A$) |
| | Can ask for confirmation of understanding during a live discussion or presentation. (P) |
| 57 | Can carry out a prepared interview, checking and confirming information as necessary. ($C_A$) |
| | Can collate information from several written sources and summarise the ideas orally. ($C_A$) |
| | Can ask for advice on a wide range of subjects. (P) |
| | Can reasonably fluently relate a straightforward narrative or description as a linear sequence of points. ($C_A$) |
| | Can respond to excuses using a range of polite forms. (P) |
| 58 | Can ask a question in a different way if misunderstood. ($N2007_A$) |
| | Can report the opinions of others. (P) |
| | Can express disagreement in a manner that shows they were actively listening to the other person. (P) |
| | Can express support in a manner that shows they were actively listening to the other person. (P) |

## GSE 59–66/B2: Speaking

| | |
|---|---|
| 59 | Can deal with less common situations in a shop, post office (e.g. returning an unsatisfactory purchase). ($C_A$) |
| | Can exchange information on a wide range of topics within their field with some confidence. ($C_A$) |
| | Can describe objects, possessions and products in detail, including their characteristics and special features. (P) |
| | Can give basic technical instructions in their field of specialisation. (P) |

| | |
|---|---|
| 60 | Can justify a viewpoint on a topical issue by discussing pros and cons of various options. ($C_A$) |
| | Can correct mistakes if they have led to misunderstandings. (N2000) |
| | Can justify and sustain views clearly by providing relevant explanations and arguments. ($C_A$) |
| | Can give the advantages and disadvantages of various options on a topical issue. ($C_A$) |
| | Can pass on a detailed piece of information reliably. ($C_A$) |
| | Can take part in routine formal discussions conducted in clear standard speech in which factual information is exchanged. ($C_A$) |
| | Can describe future plans and intentions in detail, giving degrees of probability. (P) |
| | Can paraphrase in simpler terms what someone else has said. (P) |
| | Can express an inference or assumption about a person's mood or emotional state. (P) |
| | Can talk about possibilities in the past with precision. (P) |
| | Can show interest and appreciation in conversation using a range of expressions. (P) |
| | Can bring relevant personal experiences into a conversation to illustrate a point. (P) |
| | Can describe an everyday consumer-related problem and request a correction or solution. (P) |
| 61 | Can use a limited number of cohesive devices with some 'jumpiness' in a long contribution. ($C_A$) |
| | Can engage in extended conversation in a clearly participatory fashion on most general topics. ($C_A$) |
| | Can respond to clearly expressed questions on a presentation they have given. ($C_A$) |
| | Can give detailed answers to questions in a face-to-face survey. (P) |
| | Can express feelings (e.g. sympathy, surprise, interest) with confidence, using a range of expressions. (P) |
| | Can use a range of language to make detailed comparisons of quantities. (P) |
| | Can show degrees of agreement using a range of language. (P) |
| 62 | Can make a note of favourite mistakes and consciously monitor speech for them. (C) |
| | Can construct a chain of reasoned argument. (C) |
| | Can describe how to do something, giving detailed instructions. (C) |
| | Can encourage discussion by inviting others to join in, say what they think, etc. ($C_A$) |
| | Can use a range of language to express degrees of enthusiasm. (P) |
| | Can recommend a course of action, giving reasons. (P) |
| | Can make a formal apology with detailed excuses or reasons. (P) |
| 63 | Can develop an argument giving reasons in support of or against a particular point of view. (N2000) |
| | Can give a clear, detailed spoken description of how to carry out a procedure. (C) |
| | Can describe the personal significance of events and experiences in detail. ($C_A$) |
| | Can accurately describe a problem with a product or piece of equipment. (P) |

| | |
|---|---|
| 64 | Can express views clearly and evaluate hypothetical proposals in informal discussions. ($C_A$) |
| | Can explain a problem and demand what action should be taken in an appropriate way. ($C_A$) |
| | Can summarise orally the plot and sequence of events in an extract from a film or play. ($C_A$) |
| | Can speculate about causes, consequences, hypothetical situations. (N2000) |
| | Can use stock phrases to gain time and keep the turn whilst formulating what to say. ($C_A$) |
| | Can plan what is to be said and the means to say it, considering the effect on the recipient. ($C_A$) |
| | Can make spontaneous announcements clearly and fluently. ($C_A$) |
| | Can fluently substitute an equivalent term for a word they can't recall. ($C_A$) |
| | Can compare and contrast situations in some detail and speculate about the reasons for the current situation. (P) |
| 65 | Can manage discussion on familiar topics confirming comprehension, inviting others in, etc. ($C_A$) |
| | Can describe goals using a range of expressions. (P) |
| | Can express opinions about news stories using a wide range of everyday language. (P) |
| | Can use a range of language to express degrees of reluctance. (P) |
| | Can use intonation to indicate various degrees of certainty during a discussion. (P) |
| 66 | Can summarise a wide range of texts, discussing contrasting points and main themes. ($C_A$) |
| | Can develop a clear argument with supporting subsidiary points and relevant examples. ($C_A$) |
| | Can give clear, detailed descriptions on a wide range of familiar subjects. ($C_A$) |
| | Can contribute to a conversation fluently and naturally, provided the topic is not too abstract or complex. (P) |
| | Can develop an argument well enough to be followed without difficulty most of the time. (C) |
| | Can give advice on a wide range of subjects. (P) |
| | Can outline an issue or problem clearly. ($C_A$) |

**APPENDIX B**
**SPEAKING ASSESSMENT MARKSHEET**

| Components | Descriptors | Part1 | Part2 | Part3 |
|---|---|---|---|---|
| **Pronunciation** | **EXCELLENT ---** Uses a full range of pronunciation features with precision. It is effortless to understand. | **4** | **4** | **4** |
| | **GOOD ---** Uses a wide range of pronunciation features, with only occasional lapses. It is easy to understand throughout; L1 accent has minimal effect on intelligibility. | **3** | **3** | **3** |
| | **FAIR ---** Uses a limited range of pronunciation features. Mispronunciations are frequent and cause some difficulty for the listener. | **2** | **2** | **2** |
| | **POOR ---** Speech is often unintelligible. | **1** | **1** | **1** |
| **SCORES =** | | | | |
| **TOTAL SCORE =** | | | | |

| | | | | |
|---|---|---|---|---|
| **Fluency** | **EXCELLENT ---** Speech is effortless with speed that approaches that of a native speaker. | **4** | **4** | **4** |
| | **GOOD ---** Speaks smoothly with little hesitation. | **3** | **3** | **3** |
| | **FAIR ---** Speech is slow and often hesitant and jerky. Sentences may be left uncompleted, but speaker is able to continue, however haltingly. | **2** | **2** | **2** |
| | **POOR ---** Noticeable pausing, false starts, reformulations and repetition. Difficult for a listener to perceive continuity in utterances. | **1** | **1** | **1** |
| **SCORES =** | | | | |
| **TOTAL SCORE =** | | | | |

| | | | | |
|---|---|---|---|---|
| **Grammar range and accuracy** | **EXCELLENT ---** Very strong command of grammatical structure and some evidence of difficult, complex patterns and idioms. Makes infrequent errors that do not impede comprehension. | **4** | **4** | **4** |
| | **GOOD ---** Good command of grammatical structures but with imperfect control of some patterns. Less evidence of complex patterns and idioms. Limited number of errors that are not serious and do not impede comprehension. | **3** | **3** | **3** |
| | **FAIR ---** Fair control of most basic syntactic patterns. Speaker always conveys meaning in simple sentences. Some important grammatical patterns are uncontrolled and errors may occasionally impede comprehension. | **2** | **2** | **2** |
| | **POOR ---** Any accuracy is limited to set or memorized expressions; limited control of even basic syntactic patterns. Frequent errors impede comprehension. | **1** | **1** | **1** |
| **SCORES =** | | | | |
| **TOTAL SCORE =** | | | | |

| | | | | |
|---|---|---|---|---|
| **Adequacy of vocabulary for purpose** | **EXCELLENT ---** Very good range of vocabulary with evidence of sophistication and native-like expression. Strong command of idiomatic expressions. | **4** | **4** | **4** |
| | **GOOD ---** Almost no inadequacies in vocabulary for the task. Only rare inappropriacies and/or circumlocution. | **3** | **3** | **3** |
| | **FAIR ---** Frequent inadequacies in vocabulary for the task. Perhaps frequent lexical inappropriacies and/or circumlocution. | **2** | **2** | **2** |
| | **POOR ---** Vocabulary inadequate even for the most basic parts of the intended communication. | **1** | **1** | **1** |
| **SCORES =** | | | | |
| **TOTAL SCORE =** | | | | |

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

| | | | | |
|---|---|---|---|---|
| **For each component** | • No communication possible.<br>• No rateable language. | **0** | **0** | **0** |

| | |
|---|---|
| **TOTAL SCORE FOR ALL PARTS =** | |
| **TOTAL SCORE FOR ALL PARTS / 3 =** | |
| **RESULT X 5 =** | |

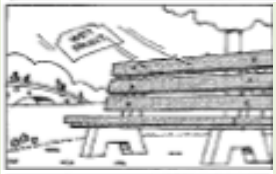
## APPENDIX C

### Overview of the test tasks
### (Before pilot testing)

### Test tasks

| Sections | Num. of quest. | Skill focus | Level | Task description | Time to plan | Time for response | Rating criteria |
|---|---|---|---|---|---|---|---|
| **Warm- up** | 3 | Personal information | A1-A2(+) 22/42 | Respond to three questions on personal topics. | No | 30 sec. for each response. | Not scored. |
| **Part 1** | 3 | Describe, express opinion, and compare with own situation | B1(+) 51/58 | Describe a picture and answer the next two questions on a concrete and familiar topic related to the photo. | No | 45 sec. for each response. | Separate task based holistic scales are used for each task. Performance descriptors describe the expected performance at each score band.<br><br>The following aspects of performance are addressed:<br><br>1) grammatical range and accuracy<br>2) lexical range and accuracy<br>3) pronunciation<br>4) fluency<br>5) cohesion and coherence. |
| **Part 2** | 3 | Describe, compare, and speculate | B1(+) 51/58 | Describe two contrasting pictures and answer two additional questions related to the topic. | No | 45 sec. for each response. | |
| **Part 3** | 3 | Narrate, speculate, and provide reasons and explanations | B2 59/66 | Narrate the sequences of events in the video and answer two additional questions related to the topic. | No | 45 sec. for each response. | |
| **Part 4** | 3 | Discuss personal experience and opinion on an abstract topic. | B2 59/66 | Integrate responses to a set of 3 questions related to a more abstract topic. | 1 minute | 2 minutes | |

**Sample questions**
**(Before pilot testing)**

| Sections | Tasks | Questions | Descriptions |
|---|---|---|---|
| **Warm-up**<br><br>**(30 seconds per question)** | Giving personal information | 1. How are you today?<br><br>2. Can you tell me where are you from?<br><br>3. What is your date of birth? | Here are some typical questions from warm-up section. Notice that all require personal information about test takers. These questions should be **easy to answer.** |
| **Part 1**<br><br>**(45 seconds per question)** | Description of a picture, giving an opinion and comparison with own situation | 1. Describe this picture.<br><br>2. Why is it important to celebrate national holidays?<br><br>3. Tell me about an important national holiday in your home country.<br><br>  | Notice that the first question is always to describe the picture.<br><br>The second question is always to **give/express/justify an opinion.**<br><br>The third question is always to **compare with own situation/experiences.** |
| **Part 2**<br><br>**(45 seconds per question)** | Description of two pictures, comparison and speculation | 1. Tell me what you see in the two pictures.<br><br>2. What would it likely to travel in these ways?<br><br>3. Which of these travelling ways would you prefer to choose?<br><br>  <br>  | Notice that the first question is always to **describe the pictures.**<br><br>The second question is always to **speculate and imagine a situation.**<br><br>The third question is always to **choose one and provide a reason for choice.** |
| **Part 3**<br><br>**(45 seconds per question)** | Narration of the story, speculation and expressing opinions. | 1. Tell me the story that the video play.<br><br>2. What could painters have done to prevent this?<br><br>3. Tell me how people usually react to the unexpected situations.<br><br>(Suppose that there is a video here in which these three pictures play in sequence)<br><br>  <br>  <br>  | Notice that the first question is always to **narrate the story in the video.**<br><br>The second question is always to **speculate about causes, consequences and hypothetical situations.**<br><br>The third question is always to **express opinions/attitudes and comment on issues related to the topic in the video.** |

| | | | | |
|---|---|---|---|---|
| **Part 4**<br><br>**(1 minute preparation, then 2 minutes to answer all 3 questions)** | Discussion of personal abstract ideas | 1. Tell me about one of the important moment in your life.<br>2. How did it affect your life?<br>3. Do hard times encourage people do the best? |  | The first question is always about **personal experiences and recall.**<br>The second question is always about **emotional responses/ feelings/opinions.**<br>The third question is always about **justifying a viewpoint.** |

**Overview of the test task specifications**
**(Before pilot testing)**

| Features of the Task | | | | |
|---|---|---|---|---|
| **Part** | Warm-up | | | |
| **Skill focus** | Introducing yourself and talking about familiar topics. | | | |
| **Task level (GSE)** | 22–29/A1; 30–35/A2; 36–42/A2 (+) | | | |
| **Task description** | The task is designed to elicit short responses on familiar and concrete topics.<br>There are three questions in this part.<br>Each question is presented separately.<br>Candidates record their responses before the next question is asked. | | | |
| **Instructions to candidates** | I will ask you three questions about yourself.<br>You will have 30 seconds to reply to each question. | | | |
| **Presentation of rubric** | Aural | Written | Other non-verbal | |
| | | | Photo | Video |
| **Planning time** | None | | | |
| **Functions** | Providing personal information | | | |
| | Explaining opinions /preferences | | | |
| | Elaborating | | | |
| | Justifying opinions | | | |
| | Comparing | | | |
| | Speculating | | | |
| | Staging | | | |
| | Describing | | | |
| | Summarizing | | | |
| | Suggesting | | | |
| | Expressing preferences | | | |

| Features of Input / Prompt | | | | |
|---|---|---|---|---|
| **Description** | 3 short questions on familiar personal topics. | | | |
| **Length of questions** | Maximum of 12 words per sentence | | | |
| **Lexical level** | | | | |
| **Grammatical level** | | | | |
| **Nature of information** | Only concrete | Mostly concrete | Fairly abstract | Mainly abstract |
| **Relevant domain** | | | | |
| **Topic** | Appropriate questions will be about familiar, everyday topics that candidates can respond to from direct, personal knowledge and experience.<br>The topics will reflect the kind of questions likely to be asked in interaction in the personal domain. | | | |

| Features of the Expected Response | |
|---|---|
| **Description** | Short responses to 3 questions at the sentence / clause level. |
| **Length of response** | Up to 30 seconds per question.<br>Adequate responses will extend beyond word/phrase level. |
| **Lexis/Grammar** | |
| **Rating Scale for task** | No scoring takes place since this part is designed to put candidates at ease, and to allow them to make the transition to speaking in the target language and to become |

| | |
|---|---|
| | comfortable with the test situation. |

| Features of the Task | | | | |
|---|---|---|---|---|
| **Part I** | Task 1 | | | |
| **Skill focus** | Describing and expressing opinions.<br>Giving brief reasons and explanations for opinions.<br>Comparing with own situation. | | | |
| **Task level (GSE)** | 51–58/B1 (+) | | | |
| **Task description** | The task is designed to elicit responses related to one picture prompt.<br>There are three questions in this part.<br>The first question asks the candidate to describe a photograph.<br>The candidate then responds to two questions related to a concrete and familiar topic represented in the photo. | | | |
| **Instructions to candidates** | I will ask you to describe a picture. Then I will ask you two questions about it.<br>You will have 45 seconds for each response. | | | |
| **Presentation of rubric** | Aural | Written | Other non-verbal | |
| | | | 1 Photo | Video |
| **Planning time** | None | | | |
| **Functions** | Providing personal information | | | |
| | Explaining opinions /preferences | | | |
| | Elaborating | | | |
| | Justifying opinions | | | |
| | Comparing | | | |
| | Speculating | | | |
| | Staging | | | |
| | Describing | | | |
| | Summarizing | | | |
| | Suggesting | | | |
| | Expressing preferences | | | |

| Features of Input / Prompt | | | | |
|---|---|---|---|---|
| **Description** | The recorded prompt asks 3 short questions related to the photograph:<br>1) describe the picture; 2) talk about an aspect of the photo relevant to the candidate's own context and experience; 3) elaborate by talking about the same topic in more general terms and provide an opinion with reasons and justification. | | | |
| **Length of questions** | Maximum of 15 words per questions | | | |
| **Lexical level** | | | | |
| **Grammatical level** | | | | |
| **Nature of information** | Only concrete | Mostly concrete | Fairly abstract | Mainly abstract |
| **Relevant domain** | | | | |
| **Topic** | The photograph will show several people engaged in an everyday, familiar activity. Appropriate questions will be about the activity and expand from asking the candidate to talk about similar activities in their own context to giving their opinions on the topic from a more general level. | | | |
| **Features of the Expected Response** | | | | |

| | |
|---|---|
| **Description** | |
| **Length of response** | Up to 45 seconds per question. Adequate responses will be beyond the single clause/sentence level. |
| **Lexis/Grammar** | |
| **Rating Scale for task** | |

| **Features of the Task** | | | | |
|---|---|---|---|---|
| **Part II** | Task 2 | | | |
| **Skill focus** | Describing familiar subjects in detail, including their characteristics and special features.<br>Comparing and contrasting situations in some detail.<br>Giving brief reasons and explanations for advantages and disadvantages.<br>Explaining the main points in an idea or problem with reasonable precision. | | | |
| **Task level (GSE)** | GSE 51–58/B1 (+) | | | |
| **Task description** | The task is designed to elicit responses related to two photos.<br>There are three questions in this part.<br>The first question asks the candidate to describe photos.<br>The candidate then will be asked to contrast and compare aspects of the photos, give opinions, and provide reasons and explanations. | | | |
| **Instructions to candidates** | I will ask you to compare two pictures and I will ask you two questions about them. You will have 45 seconds for each response. | | | |
| **Presentation of rubric** | Aural | Written | Other non-verbal | |
| | | | 2 Photos | Video |
| **Planning time** | None | | | |
| **Functions targeted** | Providing personal information | | | |
| | Explaining opinions /preferences | | | |
| | Elaborating | | | |
| | Justifying opinions | | | |
| | Comparing | | | |
| | Speculating | | | |
| | Staging | | | |
| | Describing | | | |
| | Summarizing | | | |
| | Suggesting | | | |
| | Expressing preferences | | | |
| **Features of Input / Prompt** | | | | |
| **Description** | The recorded prompt asks 3 short questions related to the photographs: 1) describe both pictures; 2) contrast and compare some aspect of the pictures; 3) provide an opinion and/or express a preference in relation to the aspects already elaborated. | | | |
| **Length of questions** | Maximum of 15 words per questions | | | |
| **Lexical level** | | | | |

| | | | | |
|---|---|---|---|---|
| **Grammatical level** | | | | |
| **Nature of information** | Only concrete | Mostly concrete | Fairly abstract | Mainly abstract |
| **Relevant domain** | | | | |
| **Topic** | The photographs will show activities/and or scenes which can be compared and contrasted. The second question will focus on some aspect of the activities/scenes open to contrast and comparison, and the third question will extend the task by asking the candidate to express an opinion and/or preference in relation to some aspect of the photos. | | | |
| **Features of the Expected Response** | | | | |
| **Description** | | | | |
| **Length of response** | Up to 45 seconds per question. Adequate responses will be beyond the single clause/sentence level. | | | |
| **Lexis/Grammar** | | | | |
| **Rating Scale for task** | | | | |

| | | | | |
|---|---|---|---|---|
| **Features of the Task** | | | | |
| **Part III** | Task 3 | | | |
| **Skill focus** | Relating a straightforward narrative as a linear sequence of points.<br>Speculating about causes, consequences, and hypothetical situations.<br>Expressing opinions and attitudes and commenting on issues. | | | |
| **Task level (GSE)** | GSE 59–66/B2 | | | |
| **Task description** | The task is designed to elicit responses related to a video prompt in which there is a series of three pictures.<br>There are three questions in this part.<br>The first question asks the candidate to narrate the sequence of actions in the video.<br>The candidate then will be asked to speculate about the possible causes, consequences, and hypothetical situations presented in the video.<br>The candidate will be asked to express opinions and elaborate on the topic. | | | |
| **Instructions to candidates** | You will look at a series of three pictures in the video. I will ask you to narrate the story that the pictures show. Then I will ask you two questions about it. You will have 45 seconds for each question. | | | |
| **Presentation of rubric** | Aural | Written | Other non-verbal | |
| | | | Photo | Video |
| **Planning time** | None | | | |
| **Functions targeted** | Providing personal information | | | |
| | Explaining opinions /preferences | | | |
| | Elaborating | | | |
| | Justifying opinions | | | |
| | Comparing | | | |
| | Speculating | | | |
| | Staging | | | |
| | Describing | | | |

| | |
|---|---|
| | Summarizing |
| | Suggesting |
| | Expressing preferences |

| Features of Input / Prompt | | | | |
|---|---|---|---|---|
| **Description** | The recorded prompt asks 3 short questions related to the video: 1) narrate the sequence of actions in the video; 2) speculate about an aspect of the video; 3) elaborate by talking about the same topic in more general terms and provide an opinion with reasons and justification. | | | |
| **Length of questions** | Maximum of 15 words per questions | | | |
| **Lexical level** | | | | |
| **Grammatical level** | | | | |
| **Nature of information** | Only concrete | Mostly concrete | Fairly abstract | Mainly abstract |
| **Relevant domain** | | | | |
| **Topic** | The video will show activities/and or scenes which can be narrated as a linear sequence of points. The second question will focus on some aspect of the activities/scenes in the video open to speculation and the third question will extend the task by asking the candidate to express an opinion and/or preference in relation to some aspect of the video. | | | |

| Features of the Expected Response | |
|---|---|
| **Description** | |
| **Length of response** | Up to 45 seconds per question. Adequate responses will be beyond the single clause/sentence level. |
| **Lexis/Grammar** | |
| **Rating Scale for task** | |

| Features of the Task | | | | |
|---|---|---|---|---|
| **Part IV** | Task 4 | | | |
| **Skill focus** | Bringing relevant personal experiences to illustrate an abstract point.<br>Expressing opinions, feelings or attitudes.<br>Justifying a viewpoint on an abstract issue by discussing pros and cons of options.<br>Showing degrees of agreement. | | | |
| **Task level (GSE)** | GSE 59–66/B2 | | | |
| **Task description** | The task is designed to elicit responses related to more abstract topic.<br>There is a set of three questions in this part.<br>The candidate speaks for two minutes after the set of questions are asked. | | | |
| **Instructions to candidates** | I will show you a picture and ask you three questions. You will have one minute to think about your answers before you start speaking. You will have two minutes to answer all three questions. | | | |
| **Presentation of rubric** | Aural | Written | Other non-verbal | |
| | | | 1 Photo | Video |
| **Planning time** | 1 minute | | | |
| **Functions targeted** | Providing personal information | | | |
| | Explaining opinions /preferences | | | |

| | | | | |
|---|---|---|---|---|
| | Elaborating | | | |
| | Justifying opinions | | | |
| | Comparing | | | |
| | Speculating | | | |
| | Staging | | | |
| | Describing | | | |
| | Summarizing | | | |
| | Suggesting | | | |
| | Expressing preferences | | | |
| **Features of Input / Prompt** | | | | |
| **Description** | The recorded prompt asks 3 short questions: 1) asks for a description of personal experience in relation to an abstract topic; 2) asks for elaboration on the candidate's impression/opinion in relation to the topic. 3) asks for a more objective discussion of the topic from the perspective of wider relevance to society/people in general.<br>A photograph is provided for extra contextualisation but is not referred to in the questions. | | | |
| **Length of questions** | Maximum of 20 words per questions | | | |
| | | | | |
| **Grammatical level** | | | | |
| **Nature of information** | Only concrete | Mostly concrete | Fairly abstract | Mainly abstract |
| **Relevant domain** | | | | |
| **Topic** | | | | |
| **Features of the Expected Response** | | | | |
| **Description** | | | | |
| **Length of response** | Up to 2 minutes for the entire long turn. | | | |
| **Lexis/Grammar** | | | | |
| **Rating Scale for task** | | | | |

## APPENDIX D

### Overview of the test tasks
### (After pilot testing)

| Sections | Num. of quest. | Skill focus | Level | Task description | Time to plan | Time for response | Rating criteria |
|---|---|---|---|---|---|---|---|
| **Warm- up** | 1 | Giving personal information | A1-A2(+) 22/42 | Respond to three questions on personal topics. | No | 30 sec. for each response | Not scored. |
| **Part 1** | 1 | Description of a picture and giving an opinion | B1(+) 51/58 | Describe a picture and answer the question on a concrete and familiar topic related to the picture. | 15 sec. | 90 sec. | |
| **Part 2** | 1 | Providing reasons and contrasting & comparison | B1(+) 51/58 | Compare pictures and provide reasons for choice. | 15 sec. | 90 sec. | |
| **Part 3** | 3 | Discussion of personal experience and opinions on abstract ideas | B2 59/66 | Integrate responses to a set of three questions related to a more abstract topic. | 1 minute | 3 minutes | |

**Sample questions**
**(After pilot testing)**

| Sections | Tasks | Questions | Descriptions |
|---|---|---|---|
| **Warm-up** | Giving personal information | 1. Please introduce yourself. | Here is a typical questions from warm-up section. Notice that it requires personal information about test takers. This question should be easy to answer. |
| **Part 1** | Description of a picture and giving an opinion | Describe this picture and talk about why it is important to celebrate national holidays.  | Notice that the first part is always to describe the picture.<br><br>The second part is always to give/express/justify an opinion. |
| **Part 2** | Comparison and providing reasons. | Look at these two pictures and talk about which of these travelling ways you would prefer to choose.   | Notice that this question is always to choose one and provide a reason for choice. |
| **Part 3** | Discussion of personal experience and opinions on abstract ideas | Talk about one of the turning points in your life.  <br><br>How did it affect your life?<br><br>Do hard times encourage people do the best? | The first one is always about personal experiences and recall.<br><br>The second one is always about emotional responses/ feelings/opinions.<br><br>The third one is always about justifying a viewpoint. |

## APPENDIX E

### Overview of the test task specifications
### (After pilot testing)

<table>
<tr><th colspan="5">Features of the Task</th></tr>
<tr><td>Part</td><td colspan="4">Warm-up</td></tr>
<tr><td>Skill focus</td><td colspan="4">Introducing yourself and talking about familiar topics.</td></tr>
<tr><td>Task level (GSE)</td><td colspan="4">22–29/A1; 30–35/A2; 36–42/A2 (+)</td></tr>
<tr><td>Task description</td><td colspan="4">The task is designed to elicit short responses on familiar and concrete topics.<br>There is one question in this part.<br>Candidates record their responses before the next question is asked.</td></tr>
<tr><td>Instructions to candidates</td><td colspan="4">I will ask you to introduce yourself.<br>You will have 30 seconds to reply.</td></tr>
<tr><td rowspan="2">Presentation of rubric</td><td rowspan="2">Aural</td><td rowspan="2">Written</td><td colspan="2">Other non-verbal</td></tr>
<tr><td>Photo</td><td>Video</td></tr>
<tr><td>Planning time</td><td colspan="4">None</td></tr>
<tr><td rowspan="12">Functions</td><td colspan="4">Providing personal information</td></tr>
<tr><td colspan="4">Explaining opinions /preferences</td></tr>
<tr><td colspan="4">Elaborating</td></tr>
<tr><td colspan="4">Justifying opinions</td></tr>
<tr><td colspan="4">Comparing</td></tr>
<tr><td colspan="4">Speculating</td></tr>
<tr><td colspan="4">Staging</td></tr>
<tr><td colspan="4">Describing</td></tr>
<tr><td colspan="4">Summarizing</td></tr>
<tr><td colspan="4">Suggesting</td></tr>
<tr><td colspan="4">Expressing preferences</td></tr>
<tr><th colspan="5">Features of Input / Prompt</th></tr>
<tr><td>Nature of information</td><td>Only concrete</td><td>Mostly concrete</td><td>Fairly abstract</td><td>Mainly abstract</td></tr>
<tr><th colspan="5">Features of the Expected Response</th></tr>
<tr><td>Length of response</td><td colspan="4">Up to 30 seconds per question.</td></tr>
<tr><td>Rating Scale for task</td><td colspan="4">No scoring takes place since this part is designed to put candidates at ease, and to allow them to make the transition to speaking in the target language and to become comfortable with the test situation.</td></tr>
</table>

| Features of the Task | | | | |
|---|---|---|---|---|
| **Part I** | Task 1 | | | |
| **Skill focus** | Describing and expressing opinions.<br>Giving brief reasons and explanations for opinions. | | | |
| **Task level (GSE)** | 51–58/B1 (+) | | | |
| **Task description** | The task is designed to elicit responses related to one picture prompt.<br>There is one question in this part.<br>The question asks the candidate to describe a picture and then respond to the question related to a concrete and familiar topic represented in the picture. | | | |
| **Instructions to candidates** | I will ask you to describe a picture and answer the question related to the picture.<br>You will have 15 seconds to think about your answer and 90 seconds to reply. | | | |
| **Presentation of rubric** | Aural | Written | Other non-verbal | |
| | | | 1 Photo | Video |
| **Planning time** | 15 seconds | | | |
| **Functions** | Providing personal information | | | |
| | Explaining opinions /preferences | | | |
| | Elaborating | | | |
| | Justifying opinions | | | |
| | Comparing | | | |
| | Speculating | | | |
| | Describing | | | |
| | Summarizing | | | |
| | Suggesting | | | |
| | Expressing preferences | | | |
| **Features of Input / Prompt** | | | | |
| **Description** | The recorded prompt asks a question related to the picture: 1) describe the picture and talk about an aspect of the photo in more general terms and provide an opinion with reasons and justification. | | | |
| **Nature of information** | Only concrete | Mostly concrete | Fairly abstract | Mainly abstract |
| **Features of the Expected Response** | | | | |
| **Length of response** | Up to 90 seconds. | | | |
| **Rating Scale for task** | **See Appendix B** | | | |

| Features of the Task | | | | |
|---|---|---|---|---|
| **Part II** | Task 2 | | | |
| **Skill focus** | Comparing and contrasting situations in some detail.<br>Giving brief reasons and explanations for advantages and disadvantages.<br>Explaining the main points in an idea or problem with reasonable precision. | | | |
| **Task level (GSE)** | GSE 51–58/B1 (+) | | | |
| **Task description** | The task is designed to elicit responses related to two pictures.<br>There is one question in this part.<br>The question asks the candidate to contrast and compare some aspects of the pictures, give opinions, and provide reasons and explanations. | | | |
| **Instructions to candidates** | I will ask you to compare two pictures, choose one and provide reasons for your choice.<br>You will have 15 seconds to think about your answer and 90 seconds to reply. | | | |
| **Presentation of rubric** | Aural | Written | Other non-verbal | |
| | | | 2 Photos | Video |
| **Planning time** | None | | | |
| **Functions targeted** | Providing personal information | | | |
| | Explaining opinions /preferences | | | |
| | Elaborating | | | |
| | Justifying opinions | | | |
| | Comparing | | | |
| | Speculating | | | |
| | Describing | | | |
| | Summarizing | | | |
| | Suggesting | | | |
| | Expressing preferences | | | |
| **Features of Input / Prompt** | | | | |
| **Description** | The recorded prompt asks a question related to the picture: 1) contrast and compare some aspects of the picture and provide an opinion and/or express a preference in relation to the aspects already elaborated. | | | |
| **Nature of information** | Only concrete | Mostly concrete | Fairly abstract | Mainly abstract |
| **Features of the Expected Response** | | | | |
| **Length of response** | Up to 90 seconds. | | | |
| **Rating Scale for task** | **See Appendix B** | | | |

| Features of the Task | | | | |
|---|---|---|---|---|
| **Part III** | Task 3 | | | |
| **Skill focus** | Bringing relevant personal experiences to illustrate an abstract point.<br>Expressing opinions, feelings or attitudes.<br>Justifying a viewpoint on an abstract issue by discussing pros and cons of options.<br>Showing degrees of agreement. | | | |
| **Task level (GSE)** | GSE 59–66/B2 | | | |
| **Task description** | The task is designed to elicit responses related to more abstract topic.<br>There is a set of three questions in this part.<br>The candidate speaks for three minutes after the set of questions are asked. | | | |
| **Instructions to candidates** | I will ask you three questions in this part. You will have one minute to think about your answers before you start speaking. You will have three minutes to answer all three questions. | | | |
| **Presentation of rubric** | Aural | Written | Other non-verbal | |
| | | | 1 Photo | Video |
| **Planning time** | 1 minute | | | |
| **Functions targeted** | Providing personal information | | | |
| | Explaining opinions /preferences | | | |
| | Elaborating | | | |
| | Justifying opinions | | | |
| | Comparing | | | |
| | Speculating | | | |
| | Describing | | | |
| | Summarizing | | | |
| | Suggesting | | | |
| | Expressing preferences | | | |
| **Features of Input / Prompt** | | | | |
| **Description** | The recorded prompt asks 3 short questions: 1) asks for a description of personal experience in relation to an abstract topic; 2) asks for elaboration on the candidate's impressions/feelings/opinions in relation to the topic; 3) asks for a more objective discussion of the topic from the perspective of wider relevance to society/people in general.<br>A photograph is provided for extra contextualisation but is not referred to in the questions. | | | |
| **Nature of information** | Only concrete | Mostly concrete | Fairly abstract | Mainly abstract |
| **Features of the Expected Response** | | | | |
| **Length of response** | Up to 3 minutes for the entire long turn. | | | |
| **Rating Scale for task** | **See Appendix B** | | | |

# APPENDIX F

## Test Item Evaluation

**Computerized Oral Proficiency Test of English as a Foreign Language (COPTEFL)**
**Test Item Evaluation – Part 1**

| | | |
|---|---|---|
| 5. Describe this picture and explain why it is important for people to have children. |  | 1. Soru anlaşılır mı?<br>EVET HAYIR<br>2. Soru amaca uygun mu?<br>EVET HAYIR<br>3. Fotoğraf net mi?<br>EVET HAYIR<br>4. Soru ve fotoğraf eşleşiyor mu?<br>EVET HAYIR<br>**Diğer görüş / düzenleme:** |
| 6. Describe this picture and explain why it is important for people to learn a new skill. |  | 1. Soru anlaşılır mı?<br>EVET HAYIR<br>2. Soru amaca uygun mu?<br>EVET HAYIR<br>3. Fotoğraf net mi?<br>EVET HAYIR<br>4. Soru ve fotoğraf eşleşiyor mu?<br>EVET HAYIR<br>**Diğer görüş / düzenleme:** |

| | | |
|---|---|---|
| 1. Look at these two pictures and explain which one you would prefer, teaching kids or adults. |  <br>  | 1. Soru anlaşılır mı?<br>EVET HAYIR<br>2. Soru amaca uygun mu?<br>EVET HAYIR<br>3. Fotoğraf net mi?<br>EVET HAYIR<br>4. Soru ve fotoğraf eşleşiyor mu?<br>EVET HAYIR<br>**Diğer görüş / düzenleme:** |

| | | |
|---|---|---|
| 2.<br><br>1. Talk about one of the challenges in your life.<br>2. How did it affect your life?<br>3. How do the difficulties people face contribute to their lives? |  | 1. Soru anlaşılır mı?<br>EVET HAYIR<br>2. Soru amaca uygun mu?<br>EVET HAYIR<br>3. Fotoğraf net mi?<br>EVET HAYIR<br>4. Soru ve fotoğraf eşleşiyor mu?<br>EVET HAYIR<br>**Diğer görüş / düzenleme:** |

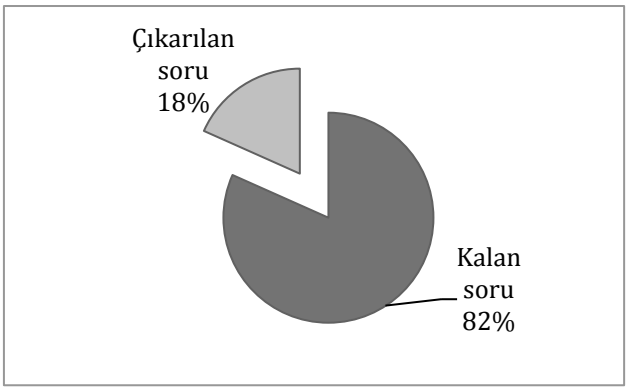
## APPENDIX G

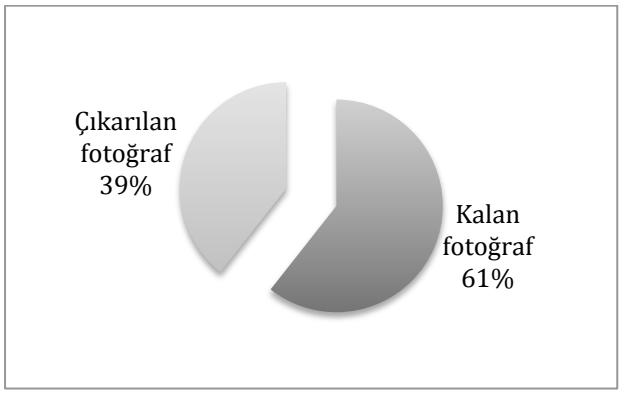
## COPTEFL Soru Düzenleme Raporu

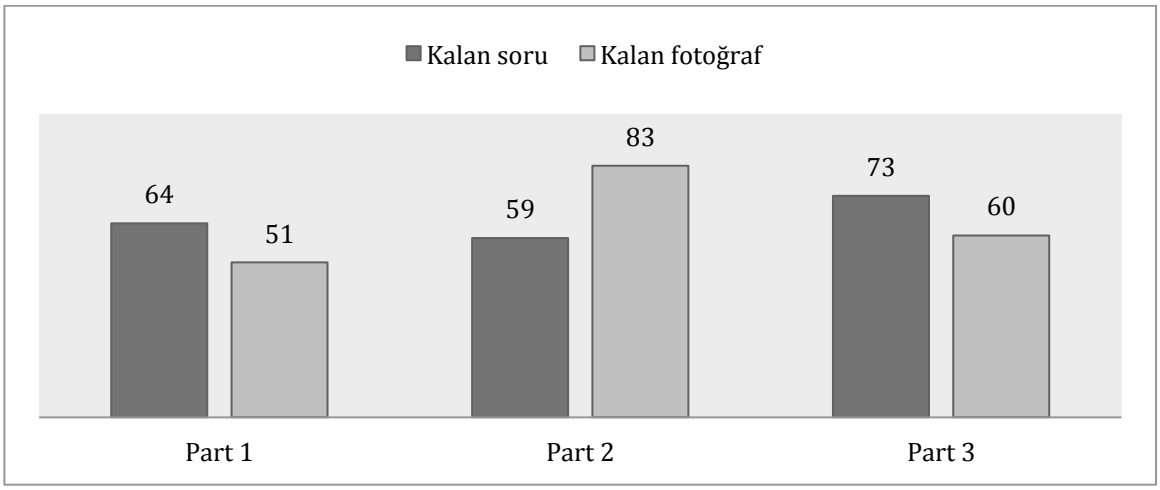
**Tablo 1: Soru sayıları**

| | TOPLAMDA | PARTLARA GÖRE | | |
|---|---|---|---|---|
| | Toplam | Part 1 | Part 2 | Part 3 |
| **Hedeflenen Soru Sayısı*** | 240 | 80 | 80 | 80 |
| **Düzenleme sonrası kalan toplam soru sayısı** | 196 | 64 | 59 | 73 |
| **Hedeflenen Fotoğraf Sayısı*** | 320 | 80 | 160 | 80 |
| **Düzenleme sonrası kalan toplam fotoğraf sayısı** | 194 | 51 | 83 | 60 |

*Hedeflenen soru sayısı: 240 ( 1 kişi = 30 soru; 8 kişi = 240 soru)
*Hedeflenen fotoğraf sayısı: 320 ( 1 set = 4 foto.; 10 set : 40; 40 X 8 = 320 fotoğraf)




**1. Editing kriterleri:**

1. Soru anlaşılır mı? (Wording açısından)
2. Soru amaca uygun mu? (see COPTEFL test task specifications tables which were given for item writing)
3. Fotoğraf net mi?

4. Soru ve fotoğraf eşleşiyor mu?

**2. Yapılan genel düzenlemeler:**

1. Part 1 ve Part 2'deki "talk about" "explain" ile değiştirildi.

2. Part 2'deki "would prefer/rather ..to/Ving" yapıları tek bir yapı olarak "would prefer Ving" şeklinde düzenlendi.

3. Part 3'deki 2. soru "how did it affect your life?" yerine "affect" dışında yeni bir soru oluşturulmuşsa "affect" li olacak şekilde değiştirildi. Örneğin;
   (1) Talk about one of the turning points in your life.
   (2) ~~How old were you at that time~~?- How did it affect your life?
   (3) Do hard times encourage people do the best?

4. Sorular task specifications şablonundaki hedefleri karşılamıyorsa çıkarıldı. Part 3'deki sorunun biraz daha üst düzey düşünceye yönlendirmesi bekleniyor. Buna yönlendirmeyen sorular çıkarıldı.

5. İçerik olarak tekrar eden sorulardan birer adet kalacak şekilde düzenleme yapıldı. Çıkarılan soruların fotoğrafları da çıkarılmış oldu. Tekrar eden sorular:
   (a) eating at home or outside,
   (b) working at office or outside,
   (c) using internet or library,
   (d) living in a flat or a house,
   (e) prefering medicine or herbs,
   (f) watching movie at home or cinema,
   (g) which cars,
   (h) which pets,
   (i) the importance of school,
   (j) the importance of learning another language,
   (k) the importance of helping others,
   (l) the importance of healthy foods,
   (m) the importance of doing exercises,
   (n) anger – how to control.

   (Birbirlerinin sorularını kontrol edenlerden de aynı soruyu soranlar olduğu görüldü)

6. İçerisinde yazı bulunan fotoğraflar çıkarıldı.

7. Fotoğraflar sınavda sorulacak boyuta getirildiğinde görselin anlaşılması zorlaşıyorsa çıkarıldı. Çıkarılan fotoğrafların yerine görsel bulundu, böylece kalan sorulara fotoğraf eklenmiş oldu. Örneğin; Part 1: 64 soru; 51 fotoğraf ; 13 adet görsel eksik kalan kısımlara yerleştirildi.

# APPENDIX H

# ETİK KURUL İZNİ

| Evrak Kayıt Tarihi: 27.03.2017 | Protokol No: 36397 | Tarih: 24.04.2017 |
|---|---|---|

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

| | |
|---|---|
| **ÇALIŞMANIN TÜRÜ:** | BAP Projesi-Doktora Tez Çalışması |
| **KONU:** | Eğitim Bilimleri |
| **BAŞLIK:** | Developing And Validating A Computerized Oral Proficiency Test of English As A Foreign Language (COPT): Turkish Context"<br>Bilgisayar Tabanlı İngilizce Konuşma Yeterlik Testi Geliştirme ve Geçerliğini Sağlama: Türkiye Ortamı |
| **~~PROJE~~/TEZ YÜRÜTÜCÜSÜ:** | Doç. Dr. Belgin AYDIN |
| **TEZ YAZARI:** | Cemre İŞLER |
| **ALT KOMİSYON GÖRÜŞÜ:** | - |
| **KARAR:** | Olumlu |

**Prof.Dr. Coşkun BAYRAK**
*(Başkan-Eğitim Fak.)*

| | |
|---|---|
| **Prof.Dr. T. Volkan YÜZER**<br>*(Başkan Yardımcısı-Açıköğretim Fak.)* | **Prof.Dr. Esra CEYHAN**<br>*(Eğitim Fak.)* |
| **Prof.Dr. Münevver ÇAKI**<br>*(Güzel Sanatlar Fak.)* | **Prof.Dr. M. Erkan ÜYÜMEZ**<br>*(İkt. ve İdari Bil. Fak.)* |
| **Prof.Dr. Handan DEVECİ**<br>*(Eğitim Fak.)* | **Prof.Dr. Emel ŞIKLAR**<br>*(İkt. ve İdari Bil. Fak.)* |