DEVELOPMENT OF A SPEAKING TEST

FOR SECOND LANGUAGE LEARNERS OF TURKISH

TALİP GÜLLE

BOĞAZİÇİ UNIVERSITY

2015

DEVELOPMENT OF A SPEAKING TEST
FOR SECOND LANGUAGE LEARNERS OF TURKISH

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfilment of the requirements for the degree of

Master of Arts

in

English Language Education

by

Talip Gülle

Boğaziçi University

2015

Development of a Speaking Test for Second Language Learners of Turkish

The thesis of Talip Gülle

has been approved by:

Assoc. Prof. Gülcan Erçetin _____

(Thesis Advisor)

Assist. Prof. Aylin Ünaldı _____

Assist. Prof. Nalan Babür _____

(External Member)

August 2015

DECLARATION OF ORIGINALITY

I, Talip Gülle, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature.....

Date .....19.08.2015.....

ABSTRACT

Development of a Speaking Test for Second Language Learners of Turkish

This study reports the findings of an initial investigation into a speaking test for learners of Turkish as a second language (L2) that is in the development process. The test aims at assessing the speaking proficiency levels of incoming exchange and Erasmus students at Boğaziçi University. For this purpose, six tasks were developed based on the Common European Framework of Reference for Languages (Council of Europe, 2001) level descriptors. These tasks were administered to a 24-person sample of L2 learners of Turkish enrolled in intermediate and advanced level Turkish language classes at Boğaziçi University. The study focused on the cognitive demands, contextual features, and "scoring validity" (Weir, 2005) of the tasks. The cognitive task demands were analyzed based on Field's (2011) cognitive processing framework for speaking, which is based on the Levelt (1989) model of speech production. The discussion of contextual task characteristics was built upon the aspects of context validity outlined by Weir (2005). Generalizability theory analyses were conducted to investigate the dependability of the scores and to examine the effect of different numbers of tasks and raters on score dependability. The findings from these qualitative and quantitative analyses provide empirical evidence into the validity and dependability of the test and inform the further refinement of the test tasks.

ÖZET

İkinci Dil olarak Türkçe Konuşma Testi Geliştirilmesi

Bu araştırma, halen geliştirilmekte olan bir ikinci dil olarak Türkçe konuşma testinin geçerlilik ve güvenilirliğine yönelik bir çalışmadır. Testin amacı, Boğaziçi Üniversitesi'ne gelen değişim öğrencilerinin konuşma yeterlilik seviyelerini ölçmektir. Bu amaçla, Ortak Avrupa Dil Referans Çerçevesi (CEFR, Council of Europe, 2001) seviye tanımlayıcıları temel alınarak altı farklı konuşma ödevi hazırlandı. Bu ödevler, Boğaziçi Üniversitesi orta ve üst düzey Türkçe sınıflarına kayıtlı 24 öğrenciye uygulandı. Çalışmada hazırlanan ödevlerin bilişsel istemleri, bağlamsal ödev unsurları, ve "puanlama geçerliliği" (Weir, 2005) incelendi. Bilişsel ödev istemleri, Field'ın (2011) konuşma için bilişsel işlemleme çerçevesine dayanarak belirlenmiştir. Bağlamsal ödev unsurları, Weir'in (2005) bağlam geçerliliği boyutları yönünden incelenmiştir. Genellenebilirlik kuramı analizleri kullanılarak, puan güvenilirliği ve farklı sayıda ödev ve puanlayıcı kullanılmasının güvenilirlik katsayısı üzerindeki etkisi hesaplanmıştır. Bu nitel ve nicel analizlerden elde edilen bulgular, hazırlanan ödevlerin geçerlilik ve güvenilirliğine dair gözlemsel kanıtlar sunmanın yanı sıra ödevlerin düzeltilmesi ve iyileştirilmesine yönelik bilgiler vermektedir.

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor Assoc. Prof. Gülcan Erçetin for providing me with guidance and encouragement throughout this study. I also want to express my heartfelt gratitude to Assist. Prof. Aylin Ünaldı for her insightful comments and support, and Assist. Prof. Nalan Babür for her encouraging feedback.

Ralph Waldo Emerson says, “because all things have contributed to your advancement, you should include all things in your gratitude.” I would like to express my gratitude to Yavuz Kurt and Ecehan Sönmez for their contributions in data collection, and Coşkun İşlek and Numan Faruk Saniç for the sleepless nights we worked together. I am also very grateful to Ayşen Tuzcu, Büşra Kamaz, İrfan Kıryar, and Sena Evren for their help with the ratings. My thanks also go to my colleagues Filiz Rızaoğlu, Elifcan Öztekin, Serhat Kurt, and Sezen Bektaş for the friendly and lively office environment; and my friends Mehmet Ali Ayaz, Rasim Çömez, Tunahan Demirbaş, Hilal Demircan, Hüseyin Demirtaş, Şükrü Erdem, Saime Kara, and Ömer Ali Tetik for supporting and helping me at all times.

I owe special thanks to Dr. Gülşen Taşdelen Teker for her help with generalizability analysis; Z. Ceyda Arslan Kechriotis and Bilgen Erdem for their guidance; and Dr. Judy Monthie-Doyum for her continued support throughout my seven years in FLED.

And, I wholeheartedly thank my family for their unconditional love and support.

My joie de vivre for having you in my life defies expression.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1
1.1 Introduction to the study 1
1.2 Aims of the present study 2
1.3 Organization of the thesis 4
CHAPTER 2: LITERATURE REVIEW 5
2.1 Introduction 5
2.2 What is 'speaking'? 5
2.3 Theoretical models of communicative competence 13
2.4 Tasks for L2 speaking tests 17
2.5 Developing test specifications 24
2.6 Validation of speaking tests 25
2.7 Generalizability theory 57
CHAPTER 3: METHODOLOGY 60
3.1 Introduction 60
3.2 The aim of the study and research questions 60
3.3 Participants 62
3.4 Instruments 64
3.5 Data collection procedures 65
3.6 Scoring procedure 67
3.7 Data analysis 67
CHAPTER 4: RESULTS AND DISCUSSION 73
4.1 Introduction 73
4.2 The construct of speaking in the current test 74
4.3 Test taker characteristics 78
4.4 Cognitive validity 81
4.5 Context validity 95
4.6 Scoring validity 107
CHAPTER 5: CONCLUSION 132
5.1 Summary and conclusion 132
5.2 Limitations 137
5.3 Suggestions for further research 138
APPENDIX A: TEST TASKS 140
APPENDIX B: TASK SPECIFICATIONS 163
APPENDIX C: RATING SCALES 182

APPENDIX D: LEARNER PROFILE FORM 193
APPENDIX E: TASK EVALUATION FORM 195
APPENDIX F: EXEMPLAR PERFORMANCES 196
APPENDIX G: VOCABULARY COUNT 217
APPENDIX H: SCORES ASSIGNED BY THE FOUR RATERS 219
REFERENCES 223

LIST OF TABLES

Table 1. Areas of Language Knowledge. 14
Table 2. Characteristics of Speaking Test Tasks. 21
Table 3. Test Taker Characteristics. 29
Table 4. Number of Participants Coming from Each Country 63
Table 5. Duration of the Speaking Tasks. 66
Table 6. Functional Requirements of the Test Tasks. 86
Table 7. Interactional Patterns and Timings for Tasks. 94
Table 8. Response Formats Used in The Current Test 95
Table 9. The Order of Tasks and Their Relevant CEFR Levels. 98
Table 10. Mean Scores of Female Test Takers on Form 1 Tasks. 104
Table 11. Mean Scores of the Test Takers in Form 1 110
Table 12. Mean Scores of the Test Takers in Form 2. 111
Table 13. Correlations among Raters by Scoring Dimension by Task for Form 1. 112
Table 14. Correlations among Raters by Scoring Dimension by Task for Form 2. 112
Table 15. Descriptive Statistics for the Analytic Scores. 113
Table 16. Variance Components for the p x t x r Design. (Form 1) 115
Table 17. Variance Components for the p x t x r Design. (Form 2) 115
Table 18. Changes in Phi Coefficients of the Four Analytic Scores in D Studies. (Form 1) 119
Table 19. Changes in Phi Coefficients of the Four Analytic Scores in D Studies. (Form 2) 119
Table 20. G Study Variance Components by Tasks. 120
Table 21. Changes in Phi Coefficients by Task. 121
Table 22. Average Agreement Rates between the Four Raters on Each Scoring Dimension by Task. 122
Table 23. Cross-Classification of Rating Frequencies for Raters 2 and 3 for Anna. 123
Table 24. Cross-Classification of Rating Frequencies for Raters 1 and 2 for Jonathan 123
Table 25. Correlations among the Analytic Scores. 129
Table 26. Proportion of Any Two Analytic Scores Differing by 1 or More and by 2 or More. 130

CHAPTER 1

INTRODUCTION

1.1 Introduction to the study

This study is centered on the development and validation of a test of speaking for learners of Turkish as a second language (L2). Hasselgreen (2004) writes that "it is difficult to find any clear, systematic way of actually testing our tests from start to finish" (p. 1). Taylor (2011), however, writes of important frameworks and models that can guide practical test design and validation. These include 'Evidence-Centered Design' and 'Assessment Use Argument' (p. 2). The validation arguments that are used in the present study draw upon another framework: the theoretical framework proposed by Weir (2005), which was operationalized in Shaw and Weir (2007), Khalifa and Weir (2009), and Taylor (2011) in terms of writing, reading, and speaking respectively, based on the Cambridge ESOL examinations. The current study is informed mainly by the operationalization of Weir's socio-cognitive framework for test development and validation within the assessment of L2 speaking ability in Taylor (2011). Another important source of reference in the development of the test tasks in the current study was the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001). The task contents were informed by the CEFR level descriptors. The assessment criteria, which are an important instrument for the operationalization of the construct in oral assessment, were also drawn from the CEFR. The current study investigates the cognitive demands and contextual characteristics of the tasks developed, and the dependability/reliability of the scores assigned to the test takers after the administration of the tasks to a group of 24 participants. The study builds upon the

operationalization of cognitive validity by Field (2011), context validity by Galaczi and ffrench (2011) and scoring validity by Taylor and Galaczi (2011).

The test used in the present study consists of six tasks (see Appendix A). Due to practicality reasons, half of the participants in the study took three of the tasks (Form 1) and the other half took the other three (Form 2). I will refer to the total of the six tasks when I write 'the (current) test' throughout the current text; and I will use the term 'the (two) forms' when a separation is made in the discussion between Form 1 and Form 2. These six tasks were developed through revisions following consultation with experts and pilot examination. In the initial test development process, a total of 15 tasks were developed (see Appendix A). Seven of these were selected to be included in the first pilot testing. Following this first pilot examination, two of the tasks were abandoned, and revisions were made on the other five to be included in the second pilot examination with the addition of one new task. The current study reports the findings obtained from the second pilot examination; however, references will be made to the first pilot examination as well wherever relevant.

1.2 Aims of the present study

Boğaziçi University, Turkey, hosts around 600 incoming exchange/Erasmus students from around 50 different countries every year, and the number of international students registered to Boğaziçi University as full-time undergraduate students is close to 300 (Boğaziçi University Office of International Relations). Each semester, about 150 of these students register to Turkish for Foreigners classes offered by the Department of Turkish Language and Literature. The department also offers a special course during the summer term, and students from around the world come to Boğaziçi University specifically for this course. The department offers classes at

different proficiency levels, but does not have a validated test for assessing incoming students' level of proficiency in Turkish so that they can be directed to the 'right' classes. Students enroll in the classes based on their own evaluations of their proficiency or the judgments of course instructors, which are based not on thorough assessment but the personal impressions they obtain from student e-mails or informal face-to-face conversations. The current test is being developed as part of a project that involves the development of reading, writing, and listening tests as well. The current study is on the development of the speaking section of this test.

Based on the six tasks that were developed as part of the current speaking test and the second pilot examination, the following research questions were asked:

1. What are the cognitive demands required by the tasks from the test takers?
2. What are the contextual characteristics of the tasks in terms of performance conditions and linguistic demands?
3. How reliable are the rating process and the scores assigned to the test takers?
4. How dependable are the scores assigned to the test takers?
   4.1. To what extent are the analytic scores dependable?
   4.2. How many tasks and raters would produce more dependable scores?
5. To what extent are the test takers' scores related to the scoring dimensions?

Briefly, the study deals with aspects of a priori validation through the analysis of test tasks in terms of their cognitive demands based on Field (2011) and context parameters based on Weir (2005); and a posteriori validation with a discussion of scoring validity, using both qualitative and quantitative methods. The study has done a considerable amount of investigation into the literature to inform the analysis of the test tasks, and includes a good deal of data analysis (analysis of performance samples and test taker scores). The data can, accordingly, be regarded as consisting of three

parts. Firstly, the test itself, which basically includes the tasks, task specifications, and rating scale (see Appendices A to C). Secondly, there is the recordings and transcriptions of test taker performances. And lastly, there is the scoring data obtained from four raters, used for quantitative analysis.

1.3 Organization of the thesis

The chapters to follow are organized as follows: Chapter 2 presents a review of the literature on three issues in general. These are (1) the definition of the ability to speak a language, or 'communicative competence', and the operationalization of 'speaking' in oral assessment, (2) test validity, focusing on cognitive validity, context validity, and scoring validity, and (3) generalizability theory, which was used to investigate score dependability under scoring validity. Chapter 3 introduces the methods and procedures used in the present study. The chapter provides detailed information on the aim of the study, research questions, participants, instruments, data collection procedure, scoring procedure, and analyses of data. Chapter 4 presents the analysis of the test tasks in terms of their cognitive and contextual characteristics as well as the results and discussion of the statistical analyses used to investigate score dependability. In Chapter 6, a summary of the findings, conclusions, and limitations of the current study are presented.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter aims to, first, consider in a broad sense what is meant by ‘speaking’ or ‘communicative language ability’; and to investigate what could be included in an operational definition of ‘speaking’ that will inform the development of a speaking test, and will contribute to the validation arguments of a test in various stages of test development from test purpose to scoring procedures. Second, drawing upon Weir (2005) and Taylor (2011), it provides an overview of issues about test validity, focusing primarily on cognitive validity, context validity and scoring validity (reliability). Criterion-related validity and consequential validity issues are beyond the scope of the present study. The chapter ends with a brief introduction to generalizability theory, since generalizability and decision studies (both are informed by generalizability theory) were used in the present study to investigate the dependability of test scores and the influence on score dependability of using different numbers of tasks and raters.

2.2 What is ‘speaking’?

Speaking is a complex skill and a difficult construct to define. Lado (1961, cited in Fulcher, 2003) writes that despite being the most valued language skill, “testing the ability to speak a foreign language is perhaps the least developed and the least practiced in the language testing field,” which he continues to argue is due to “a clear lack of understanding of what constitutes speaking ability and oral production” (p. 239). Twenty years later, Bachman and Palmer (1981) pointed to the complexity of assessing speaking, arguing that the measurement of oral proficiency continued to be one of the areas of most persistent difficulty in the field of language assessment.

There have been various attempts since then to understand and define this complex skill and these have contributed to our understanding of the ability to speak a language, whether it is a first language (L1) or a second language (L2).

Noting the lack of an operational definition of the construct of 'speaking' Lado's (1961) approach, which came to be known as "trait theory," (Fulcher, 2003) was to eliminate variables, such as introversion, which could confound with the construct and thus influence the interpretation of the scores. However, more recent investigations of the construct of speaking have emphasized the role of contextual factors (Bachman & Palmer, 1996; Douglas, 2000, cited in Fulcher, 2003), particularly when such factors influence the discourse produced by test takers and their scores, for example in the case of tests of speaking for specific purposes, where the construct definition, the selection of tasks and the scoring relate to a particular context (Fulcher, 2003).

Fulcher (2003) writes that construct definition includes questions like 'what is speaking?' or 'what constitutes speaking?', yet immediately points out that such a complex process as communication cannot be captured in its totality by any operational construct definition. He argues that what defines the borderline for a construct definition is usually the test purpose and the expected level of generalizability of test scores to the non-test situations that the test aims at. He emphasizes the usefulness of construct for the testing purpose. Because "we should not assume that any description, any rating scale, captures some psychological reality that exists in the language competence of all speakers for all time in all contexts," the construct in a test situation should be defined in ways that are relevant and meaningful to the test-takers, be based on the test purpose, and reply to the requirements of score users (Fulcher, 2003, p. 19). He writes:

> The construct definition we decide upon should not be evaluated by its correspondence to psychological reality, but its utility in making inferences from test scores, and its value in helping us to construct a validity argument that supports the link between a test score and what we claim it means. (p. 19)

So, Fulcher (2003) points out that any construct of speaking will be multi-faceted, and discusses various components that could be included in the construct definition of speaking in testing situations. These components include pronunciation and intonation, accuracy and fluency, strategic competence, pragmatic appropriacy, and interactional competence. The following sections will present a brief introduction to these aspects for the conceptualization and operationalization of speaking in general, and to provide some background on these sub-skills since they are included in the theoretical models of communicative competence which will be reviewed in the sections to follow and since they also contribute to the construct definition of the test in the current study, which is reflected in the rating scale.

Pronunciation: Luoma (2004) discusses pronunciation in relation to the currently debated issue of whether and, if so, how we should question the norms of the native speaker standard. She argues that the native speaker standard for foreign language pronunciation can be questioned on certain grounds. First of all, identifying a single standard as the native speaker standard is difficult. Moreover, very few second language learners succeed in achieving a native-like standard in all aspects of a language. Similarly, McNamara (1996) points to the difficulty of deciding on the form of Standard English that should be used as the target for learners and test takers. Although pronunciation is commonly included as one of the criteria on which test takers are assessed (for example, in IELTS or other Cambridge ESOL tests; or Pearson Test of English General), Fulcher (2003) argues that since for most learners pronunciation is “only a matter of general intelligibility”, the inclination is towards

assessing learners in terms of intelligibility instead of how well they conform to a standard variety (p. 25).

Accuracy and fluency: Accuracy and fluency are commonly treated as distinct, albeit related, components of speaking in both SLA and language teaching and language testing literature (Fulcher, 2003). While some examination boards assign one score for these two components, others give separate scores. In the latter practice, the accuracy of grammatical structures and lexical resources is seen as one component of assessment, and the quality and speed of speech as another. Based on Skehan and Foster (1999), Ellis (2009) defines accuracy as "the ability to avoid error in performance, possibly reflecting higher levels of control in the language as well as a conservative orientation, that is, avoidance of challenging structures that might provoke error" (p. 2). Definitions of accuracy generally relate to the ability to produce error-free utterances. The operationalization of fluency, on the other hand, has been varied. Fulcher (1996) suggests that despite being commonly utilized as criteria in speaking tests and in the language assessment literature, fluency is not defined adequately for the purposes of operationalization and therefore rating scales generally include an inadequate operationalization of the concept of fluency. Douglas (1994), for example, defined fluency as the ratio of types and tokens in the transcriptions of student speech; however, he found little relationship between the fluency scores and the language actually produced (the ratio of types and tokens). Various researchers have have commented on what makes speech (non)fluent. Chambers (1997) lists a number of features that are associated with (dys)fluency:

- the frequency of pauses rather than the length,
- the length of run (the number of syllables between pauses),
- the place of pauses in an utterance, and
- the transfer (or not) of pausing pattern from L1 to L2. (p. 541)

Fulcher (2003) offers the following phenomena as indicators of dysfluency:

- Hesitations consisting of pauses, which can be unfilled (silence) or filled (with noises like 'erm').
- Repeating syllables or words.
- Changing words.
- Correcting the use of cohesive devices, particularly pronouns.
- Beginning in such a way that the grammar predicts what comes next, but the speaker changes the structure of the utterance part way through. (p. 30)

Skehan (2009), on the other hand, writes that fluency measures examine the following:

- breakdown (dys)fluency, indexed by pausing;
- repair (dys)fluency, indexed by measures such as reformulation, repetition, false starts, and replacements; and
- speed, with measures such as syllables per minute. (p. 3-4)

Fulcher (2003) writes that "The fact that observable speech phenomena are … not 'linear' (that is, more pauses does not necessarily equal less fluency) is problematic for testing speaking, which has traditionally relied on linear scales for scoring" (p. 31). Therefore, the formulation of such lists should not be taken readily to suggest that any of the features in these lists is in a linear relationship with proficiency levels. As Fulcher suggests, not all instances of pausing, for example, will indicate low proficiency. This is related to "non-linearity in language development" (Fulcher, 2003, p. 31). This also implies that speakers with high proficiency, like those with lower proficiency, may hesitate or pause while speaking. The reason for hesitations or pauses may not always be searching for language, but for constructing or organizing upcoming ideas and utterances, which is common for proficient speakers of a language especially when the topics they talk on or the ideas that they formulate are complex. Fulcher (1996), for instance, investigated the reasons for interruptions in fluency and found the following categories to account for such interruptions: end-of-turn pauses, content planning hesitation, grammatical planning hesitation, addition of examples/counter-examples/reasons to support a point of view, expressing lexical

uncertainty, grammatical and/or lexical repair, expressing propositional uncertainty, and misunderstanding or breakdown in communication. Therefore, the operationalization of fluency in rating scales needs to tap into the causes of markers of dysfluency rather than equating, for example, each pause or false start with inability to speak fluently.

Strategic competence: In language testing literature, Canale and Swain (1980) is frequently referred to for their definition of 'strategic competence'. They define strategic competence as "verbal and nonverbal communication strategies that may be called into action to compensate for breakdowns in communication due to performance variables or to insufficient competence" (p. 30). They suggest that these strategies are used when a deficiency in grammatical or sociolinguistic competence causes communication difficulties. In Bachman's model, strategic competence is "a general ability, which enables an individual to make the most effective use of available abilities to carry out a given task" (p. 106). Bachman and Palmer's (1996) definition of strategic competence, on the other hand, includes metacognitive strategies such as "goal setting, assessment, and planning" (p. 70). Fulcher (2003) argues that only a few tests include strategy use in their assessment criteria, because it is difficult to determine strategy use by speakers and even more so to attribute purpose to these uses.

Pragmatic competence: Hymes (1971, cited in Fulcher, 2003) asserts that there are rules of use and unless the speaker had the knowledge of them, the rules of grammar would be useless. In other words, he argues that communicative language ability cannot be restricted to the knowledge of the grammatical or phonological system of a language; pragmatic appropriacy is an important consideration. Other researchers from the field of sociolinguistics have investigated what features of communication

relate to pragmatic appropriacy. Grice (1975), for example, argues that there are four overriding pragmatic factors (or 'maxims' in Grice's terms) that speakers need to conform to in their speech to communicate efficiently. These factors are: quantity (giving enough information, not too little or too much), quality (saying only what you know to be true), relation (being relevant to the topic of conversation), and manner (being clear and orderly in interaction). However, Brown and Levinson's (1987) politeness theory puts forward that speakers want to protect their 'face' (public self-image) in communicating with others, and this may cause them not to follow Grice's maxims. In addition to Grice's maxims and face protection, the social distance between the interlocutors and the degree of imposition of a speech act may influence the way speakers interact (Austin, 1975). Fulcher (2003) points out that pragmatic features may be accommodated during test design, or they may become an asset of the scoring criteria based on the test purpose.

Interactional competence: The importance of defining and operationalizing interactional competence in speaking tests was highlighted by Kramsch (1986). The term 'interactional competence' was later embraced by He and Young (1998) in Interactional Competence Theory (ICT). In ICT, the definition of interactional competence encompasses the view that speaking is a co-constructed phenomenon. The co-construction of speech has been addressed by many researchers (Galaczi, 2008; May, 2009; 2011; McNamara, 1996, 1997; Swain, 2001; Taylor & Wigglesworth, 2009). McNamara (1996) argues that "a weakness of current models is that they focus too much on the individual rather than the individual in interaction" (p. 85). He distinguishes between two uses of the term 'interactional competence', suggesting that the construct of interaction is defined in models of Communicative Language Ability (CLA) in a psychological sense and relates to a single individual,

and that a social/behavioral perspective views interaction as joint behavior between individuals. He and Young (1998) make a clear statement regarding the latter use of the term from a social perspective:

> Interactional competence is not an attribute of an individual participant, and thus we cannot say that an individual is interactionally competent; rather we talk of interactional competence as something that is jointly constructed by all participants (including an analyst if the interaction is subject to analysis). Equally, interactional competence is not a trait that is independent of the interactive practice in which it is (or is not) constituted. (p. 7)

Fulcher (2003) points out that ICT raises a number of questions for construct definition in tests of second language speaking ability, such as "how scores can be given to an individual test taker rather than pairs of test takers in paired test format" or "how the contributions of the interlocutor/rater can be taken into consideration when he or she is co-responsible for the construction of the talk" (p. 46). Although the common practice of examination boards is to rate test takers individually in interactive tasks, research suggests that raters are aware of the co-constructed nature of interactions (Galaczi, 2008; May, 2009, 2011). Fulcher (2010), on the other hand, argues that interactional competence can be seen as "a set of abilities that an individual brings to the temporally bounded interaction set within a specific context and discursive practice" (p. 112). Viewing interaction in this light may justify assigning separate scores to test takers in tasks that require interaction. There seems to be a need for further research for the operationalization of 'interaction' in tests of oral proficiency.

The features and components of 'speaking' that have been presented so far are explicated and investigated in relation to each other in an explicit format in models of communicative competence. The next section will provide an overview of these models.

2.3 Theoretical models of communicative competence

As was pointed out above, while developing speaking tests, the developers aim at being able to provide the evidence that the scores the test takers obtain in the test are relevant to a particular outside-the-test context. Luoma (2004) argues that even when the purpose of the test is general, as is the case for language proficiency tests, a reference framework is necessary to operationalize what is meant by language proficiency and in what ways the test relates to it. Because theoretical models are generally presented in an explicit format, it is possible to choose one or more of them depending on the testing situation and view the construct definition of the ability assessed in the test in the light of the selected models (Luoma, 2004).

Communicative language ability has been defined in different ways (it should be noted that they also show substantial similarities) by various researchers and a number of theoretical models have been proposed to explain what components constitute what we call 'communicative language ability'. Below I will present some of these models that can be of use in the development of a speaking test, with particular focus on the model proposed by Bachman and Palmer (1996) and the way 'communicative language ability' is conceived in Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001).

Canale and Swain (1980) proposed a multi-componential model of communicative competence. Canale and Swain's model includes sociolinguistic, discourse, and strategic competence in addition to grammatical competence. This model is credited for providing a theoretical framework for communicative language testing and teaching (Purpura, 2008). A more comprehensive model of communicative competence was proposed by Bachman (1990), and this model was refined by Bachman and Palmer (1996). Table 1 presents the areas of language

knowledge proposed by Bachman and Palmer (1996). The model presented by Bachman (1990) and Bachman and Palmer (1996) is seen as a further development of the Canale and Swain model as it distinguishes between what constitutes 'knowledge' and what constitutes a 'skill' and attempts to define the interaction among various components, as well as between these components and the language use context (Fulcher & Davidson, 2007).

Table 1. Areas of Language Knowledge.

Organizational knowledge (how utterances or sentences and texts are organized)
Grammatical knowledge (how individual utterances or sentences are organized)
Knowledge of vocabulary Knowledge of syntax Knowledge of phonology/graphology
Textual knowledge (how utterances or sentences are organized to form texts) Knowledge of cohesion Knowledge of rhetorical or conversational organization
Pragmatic knowledge (how utterances or sentences and texts are related to the communicative goals of the language user and to the features of the language use setting)
Functional knowledge (how utterances or sentences and texts are related to the communicative goals of language users)
Knowledge of ideational functions Knowledge of manipulative functions Knowledge of heuristic functions Knowledge of imaginative functions
Sociolinguistic knowledge (how utterances or sentences and texts are related to features of the language use setting)
Knowledge of dialects/varieties Knowledge of registers Knowledge of natural or idiomatic expressions Knowledge of cultural references and figures of speech

Source: Bachman and Palmer, 1996.

According to Bachman and Palmer (1996), communicative language ability is influenced by such traits of language users as their topical knowledge, affective schemata, language ability as well as some general characteristics. Language ability,

in their model, is composed of two broad areas. These are language knowledge and strategic competence. Language knowledge, then, is divided into organizational knowledge, which is comprised of grammatical knowledge and textual knowledge, and pragmatic knowledge, which is comprised of functional knowledge and sociolinguistic knowledge. Grammatical knowledge in this model includes knowledge of vocabulary, phonology, morphology, syntax, and graphology. These enable the language user to recognize and produce grammatically correct sentences (p. 68). Textual knowledge includes knowledge of conventions for combining sentences or utterances into text. In other words, textual knowledge relates to knowledge of cohesion and knowledge of rhetorical organization, or conventional organization (p. 68-69). Lastly, pragmatic knowledge enables the language user to express acceptable language functions and interpret intentions and discourse, and to create and interpret language utterances that are appropriate in a particular context of language use. These subcomponents of pragmatic knowledge are named functional knowledge and sociolinguistic knowledge, respectively (p. 69-70). Strategic competence, on the other hand, represents a set of metacognitive components, thanks to which the language user is involved in goal setting, assessment of communicative sources, and planning (p. 70).

However, in their investigation of facets of speaking proficiency, De Jong, Steinel, Florijn, Schoonen, and Hulstijn (2012) note that although both the Canale and Swain (1980) model and the Bachman and Palmer (1996) model include linguistic knowledge, they make no explicit references to linguistic processing skills, which they define as the ability to rapidly and correctly process linguistic information.

The model described in the CEFR (Council of Europe, 2001) shows substantial similarities with Bachman and Palmer's (1996) model, but it also incorporates linguistic processing. In this model, communicative competence includes three basic components: language competence, sociolinguistic competence and pragmatic competence. In the CEFR, language competence is comprised of lexical, grammatical, semantic, phonological, orthographic, and orthoepic competences. Sociolinguistic competence relates to the language elements that mark social relationships, rules of appropriate behavior and expression of people's wisdom, differences in register and dialects, and stress. Lastly, "*pragmatic competences* are concerned with the user/learner's knowledge of the principles according to which messages are a) organized, structured and arranged ('discourse competence'); b) used to perform communicative functions ('functional competence'); c) sequenced according to interactional and transactional schemata ('design competence')" (Council of Europe, p. 123). In contrast to Bachman and Palmer's model, in the CEFR, strategic competence is conceived as strategy use in the broadest sense as "a hinge between the learner's resources (competences) and what he/she can do with them (communicative activities)" (p. 25). Moreover, the CEFR explicitly refers to linguistic processing skills in terms of production as well as interaction:

> 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. (p. 158-159)

It is also argued in the CEFR that a task is likely to place heavier or lighter cognitive demands on the learner depending on his/her capacity to (a) "handle the number of steps or 'cognitive operations' involved, and their concrete or abstract nature" and (b) "attend to the processing demands of the task (amount of 'on-line thinking') and to relating different steps of the task to one another (or to combining different but related tasks)" (p. 161).

So far, attempts to define the ability to speak a language have been reviewed. These have contributed substantially to our understanding of the construct of 'speaking'. Since the construct of speaking in tests of L2 speaking ability is operationalized through test tasks, the focus of the following section will be tasks.

2.4 Tasks for L2 speaking tests

As is the case with other steps in test development, in designing speaking test tasks, one of the main concerns is the construct-related information that the scores must deliver, which means that the tasks should make the test takers show the type of skills that are specified in the construct definition and test specifications (Fulcher, 2003). In order to do this, care must be taken in the task design process to avoid two threats to construct validity proposed by Messick (1989): construct underrepresentation (failure to capture important aspects of the construct in the test) and construct-irrelevant variance (the impact upon scores of factors that are irrelevant to the construct measured in the test). Below, some of the decisions that test developers need to make and which can impact upon the construct validity of a test are presented.

One of the major considerations in developing a test of speaking proficiency is whether to use direct or semi-direct modes of speaking assessment. A direct test "requires the candidate to perform precisely the skill that we wish to measure"

(Hughes, 2003, p.17). Clark (1975) defines direct testing in relation to the assessment of speaking ability as follows:

> In direct proficiency testing, the testing format and procedure attempts to duplicate as closely as possible the setting and operation of the real-life situations in which the proficiency is normally demonstrated… A direct test of oral proficiency, in the face-to-face communication sense, would involve a test setting in which the examinee and one or more human interlocutors do, in fact, engage in communicative dialogue. (p. 10-11)

A semi-direct test, on the other hand, uses non-human elicitation methods (Clark, 1975). In the past, semi-direct tests of oral proficiency were generally tape-mediated; today, they are computer- or Internet-based. Although a number of studies comparing these two modes indicated high score equivalency (Bernstein, Van Moere & Cheng, 2010; Kenyon & Tschirner, 2000; Weir & Wu, 2006), they differ in a number of important ways. Comparing the two modes, Qian (2009) writes that while semi-direct testing reduces the need for export resources, is cost-effective and efficient, and uses standardized instructions and prompts, thus eliminating variation in examiner input, it is seen as inferior to direct testing with respect to test validity as it does not reflect typical real-life communication. Direct testing can, therefore, be argued to be more authentic, which provides validity evidence for a test; yet, it is also argued that authentic task scores lack reliability and generalizability (Tarone, 1998).

A recent study by Jeong et al. (2011), which used functional magnetic resonance imaging (fMRI), showed that L2 direct interview produced significantly greater activation in brain regions concerned with social communication than L2 semi-direct interview, and similar brain regions were activated during L1 direct, L1 semi-direct and L2 direct interviews. This study points to the superiority of direct testing in terms of validity when the test includes social communication in its construct definition. However, as Taylor (2011) points out, “choice of oral proficiency test method will invariably depend on a variety of complex factors” such

as “heavy administrative demands”, “costs” as well as “factors relating to validity and, within that, reliability” (p. 24).

Another important decision that test developers need to make at this stage is what the test takers will be doing with language since speakers’ use of language is argued to differ in situations which require different types of speech (Luoma, 2004). Brown and Yule (1983, cited in Luoma), for example, list four different types of informational talk: description, instruction, storytelling and opinion expressing/justification. Bygate (1897, cited in Luoma) presents a more detailed categorization:

- Factually oriented talk: description; narration; instruction; comparison.
- Evaluative talk: explanation, justification, prediction, decision.

The Common European Framework of Reference (CEFR) makes a distinction between functional and notional competence (Council of Europe, 2001). Functional categories are then divided into macro- and micro-functions. While macro-functions include description, narration, commentary, explanation and demonstration; micro-functions are related to individual actions that require a turn in interaction. Micro-functions include, but are not limited to, reporting, asking, expressing agreement/disagreement, suggesting, requesting, greeting, introducing, summarizing, changing the theme, signaling non-understanding, or paraphrasing.

Other decisions related to speaking test tasks include whether test takers should be tested at a time in an interview format, in pairs or in groups; whether the tasks should be integrated or independent; whether task-based or construct-based assessment should be used; what the interaction type should be; what types of tasks should be used, open-ended or structured, among many others (Luoma, 2004). Each of these decisions will have consequences for the particular testing situation. For

example, while individual testing offers certain advantages such as flexibility to adapt questions according to the performance of a test taker and providing control over the interaction, it is not economic in terms of time (Luoma, 2004). Moreover, individual testing has been subject to criticism since in this type of testing the interviewer has considerable power over the test taker (Ross & Berwick, 1992; Taylor, 2001; Young & Milanovic, 1992). An alternative to individual testing is peer-to-peer interaction through, for example, information gap tasks. Such tasks offer broader evidence for language proficiency, especially since speaking ability includes interaction skills (Luoma, 2004). However, paired tasks may not provide equal opportunity to the test takers to show their speaking skills since one test taker may dominate the other (Weir, 1993). As such, each task type has its own advantages and disadvantages. Therefore, the decision to select a particular task type over another should be informed by the testing context.

For any test, the degree of correspondence between test tasks and target language use tasks should be investigated in order to identify what "the test takers will be asked to do with language" (Luoma, 2004, p. 31). To this end, Bachman and Palmer's (1996) framework of task characteristics serves as a useful tool (See Table 2).

Bachman and Palmer (1996) propose that defining task characteristics is important in helping to show how speaking in the real word is reflected in test tasks. Similarly, Fulcher (2003) suggests that the development of such lists of task characteristics relate to the discussion of 'authenticity' of language tasks. Fulcher (2003) also argues, however, that since there have been few investigations into which task features should be included in test specifications for speaking tests or which features are more important in the description of tasks, it is a challenging undertaking

to produce comparable tasks using such categorizations. These lists also raise the question of whether it is possible to match test tasks to real-world tasks with reference to all of the characteristics included in such lists (Fulcher, 2003).

Table 2. Characteristics of Speaking Test Tasks.

- Characteristics of the setting
  - Physical characteristics
  - Participants
  - Time of task
- Characteristics of the test rubrics
  - Instructions
    - Language (native, target)
    - Channel (audio, visual)
    - Specification of procedures and tasks
  - Structure
    - Number of parts/tasks
    - Salience of parts/tasks
    - Sequence of parts/tasks
    - Relative importance of parts/tasks
    - Number of tasks/items per part
  - Time allotment
  - Scoring method
    - Criteria for correctness
    - Procedures for scoring the response
    - Explicitness of criteria and procedures
- Characteristics of the input
  - Format
    - Channel (aural, visual)
    - Form (language, non-language, both)
    - Language (native, target, both)
    - Length
    - Type (item, prompt)
    - Degree of speededness
    - Vehicle (live, reproduced, both)
  - Language input
    - Language characteristics
      - Organizational characteristics
        - Grammatical (vocabulary, syntax, phonology, graphology)
        - Textual (cohesion, rhetorical/ conversational organization)
      - Pragmatic characteristics
        - Functional (ideational, manipulative, heuristic, imaginative)
        - Sociolinguistic (dialect/variety, register, naturalness, cultural references and figurative language)
    - Topical characteristics

Source: Bachman and Palmer, 1996.

A further consideration in task-based assessment is task difficulty. Fulcher (2003) argues that the choice of a suitable range of tasks for assessment purposes and the way levels of test performance are described could be more informed if test developers had an understanding of what factors are influential in task difficulty.

Brown, Anderson, Shillcock and Yule (1984) propose that as tasks move from static to dynamic to abstract, task difficulty increases. They also suggest that task difficulty is influenced by the number of participants or elements included in the task and the relationship between them. They argue, for example, that it is more difficult to describe a story with four female characters talking to each other than a story which involves one male and one female character.

To define task difficulty, Skehan (1998) developed a model consisting of three different factors:

- Code complexity (e.g. linguistic complexity, vocabulary load);
- Cognitive complexity (e.g. cognitive familiarity such as familiarity of topic/discourse genre/task, and cognitive processing factors such as information type and organizational structure);
- Communicative stress (e.g. logistics of test performance such as time pressure, number of participants, or type of response).

Elder, Iwashita, and McNamara (2002) investigated the effect of various performance conditions on perceptions of task difficulty and on test scores in a test of speaking using Skehan's framework. They used eight different story tasks, with two different performance conditions for each story. The tasks differed in terms of perspective (telling a story as it happened to the test taker vs. telling a story form someone else's perspective), immediacy (here and now vs. there and then), adequacy (telling a story from a set of six pictures vs. telling the same story with one picture missing), and planning time (3 minutes of pre-planning vs. no planning). They found no systematic variation associated with various performance conditions, thus offering little support for Skehan's framework in the context of speaking assessment. The

study, therefore, pointed to the difficulty of making a priori estimates of task difficulty in oral proficiency assessment.

Despite attempts to develop frameworks to define task difficulty, a number of other studies as well failed to find score sensitivity to variation in task conditions (Fulcher 1996b; Fulcher & Reiter, 2003; Bachman, Lynch, & Mason, 1995; Lumley & O'Sullivan, 2005). Fulcher (1996b) used three different tasks; two were one-to-one interviews and one a group discussion. The first task required the description of a picture and then a discussion on a related topic. In the second task, the students were asked to discuss a previously encountered text. For the third task, the students were asked to discuss on the topic of education in their country following a 10-minute preparation time. The study found that the task effect on scores was so small that the scores were generalizable from one task to another. Fulcher and Reiter (2003) aimed to maximize task variance, using six tasks that varied systematically in terms of social power and degree of imposition. The study pointed to differences in the discourse produced in different tasks, but it was found that this difference did not translate into changes in the scores. Lumley and O'Sullivan (2005) report similar findings in their investigation of the effect of gender, audience, and topic in task difficulty, and point out that "we are still a long way from making accurate predictions of task difficulty" (p. 420). They also write, "Tasks are more likely to affect individuals differentially, rather than at the level of groups defined at this level," (p. 433) that is, at the level of gender. Fulcher (2003) suggests that it is not clear precisely what makes a task difficult or easy for a particular group of test takers, mainly "because 'difficulty' does not reside in the task itself, but is an interaction of tasks, conditions and test takers" (p. 67).

The sections so far have provided an overview of what constitutes the construct of ‘speaking’ and of a number of important decisions test developers need to make, which are informed by the test construct. The relationship between the test construct and the test tasks is captured in test specifications by test developers. Test specifications will be the focus of the next section.

2.5 Developing test specifications

Test specifications are the detailed written version of test developers’ ideas regarding such matters as what kinds of speaking the assessment will be aiming at, how the assessment procedure will be carried out, as well as what the scoring criteria will be, among others. The specifications contain the construct definition, detailed definitions of the tasks the test involves, the rating criteria, the explanations as to why the assessment focuses on certain constructs, and how the tasks and scoring operationalize them (Luoma, 2004). One of the most detailed lists of contents for test specifications can be found in Alderson, Clapham and Wall (1995). They indicate that the specifications should include: the test’s purpose; description of the examinees; test level; definition of construct (theoretical framework for the test); description of suitable language course or textbook; number of sections/papers; time for each section/paper; weighting for each section/paper; target language situation; text-types; text length; language skills to be tested; language elements to be tested; test tasks; test methods; rubrics; criteria for marking; descriptions of typical performance at each level; description of what candidates at each level can do in the real world; sample papers; and samples of student performance on task (p. 11-20).

Fulcher (2003) lists the types of activities that are undertaken in the process of developing test specifications, and writes that these should be documented along with the specifications themselves. These are:

- Identifying members of the design team.
- Identifying the test takers.
- Defining test purpose.
- Defining the construct.
- Designing prototype tasks that elicit speech relevant to the constructs.
- Piloting prototype tasks with small groups of learners in order to discover which tasks do not work as planned, and which can be included in the test after revision and further piloting.
- Working on initial ideas for rating scales and band descriptions as prototyping continues through iterations of pilots with small groups of students.
- Writing and revising the task administration instructions.
- Commissioning appropriate research to support design decisions.
- Making explicit any constraints in test design. (p. 117-119)

In all of these activities, from the definition of the test construct to the investigation of constraints in test design following test administration, an overarching consideration is test validity. In fact, reference has been made to validity various times in the previous sections, which indicates that validity is a concern in every step in test development. The following sections will provide an overview of issues related to the validation of speaking tests.

2.6 Validation of speaking tests

The approach and methodology used in the present study is informed by the theoretical framework for validating language examinations outlined in Weir's (2005) *Language Testing and Validation: An evidence-based approach*, and operationalized in Shaw and Weir (2007), Khalifa and Weir (2009), and Taylor (2011) as it was applied to the assessment of reading, writing, and speaking, respectively. As is proposed by Taylor (2011), this socio-cognitive framework for validating examinations can be used for "a systematic and comprehensive critical evaluation of construct definition and operationalization" (p. 8). Taylor (2011) writes that the framework is socio-cognitive in that it views the use of language as a social phenomenon rather than a purely linguistic one and the abilities of test takers are

shown in relation to the mental processing that takes place as they engage in test tasks.

Before moving on to the components of the framework, a point needs to be made from the beginning; practicality is not an immediate concern for Weir (2005):

> Our heretical view (and certainly not one shared by Exam Boards) is that practicality is simply not a necessary condition for validity. Only when sufficient evidence is available to justify interpreting test scores as an acceptable indication of the control of an underlying construct should we concern ourselves with practicality. If practicality is allowed to intrude before such evidence is available we run the risk of not assessing what we want to. The convenience of the method should not be allowed to subvert the measurement of the construct. (p. 49)

In Figure 1, Weir's framework for validating speaking tests is provided. All of the discussions which follow will essentially be built upon this framework.

Weir (2005) explains that "the arrows indicate the principal direction(s) of any hypothesized relationship: what has an effect on what, and the timeline runs from top to bottom; before the test is finalized, then administered and finally what happens after the test event" (p. 43). Taylor (2011) lists a number of questions that should be considered in applying the socio-cognitive validation framework to language tests. These are:

- How are the physical/physiological, psychological, and experiential characteristics of candidates catered for by this test?
- Are the cognitive processes required to complete the test tasks appropriate?
- Are the characteristics of the test tasks and their administration appropriate and fair to candidates who are taking them?
- How far can we depend on the scores which result from the test?
- What effects do the test and test scores have on various stakeholders?
- What external evidence is there that the test is measuring the construct of interest?

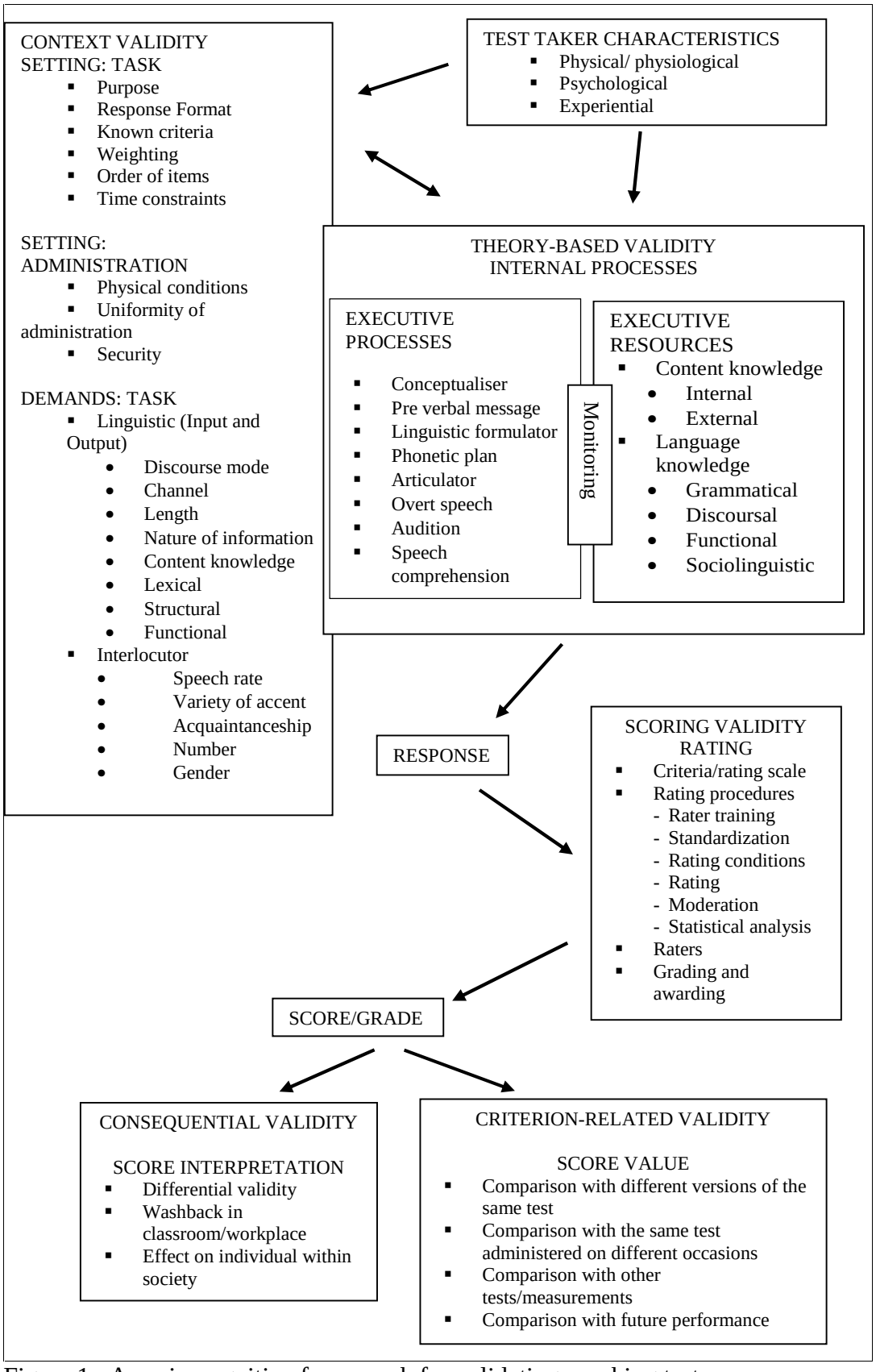


Figure 1. A socio-cognitive framework for validating speaking tests.
Source: Weir, 2005, p. 46.

The first question relates to test taker characteristics, the second to cognitive validity, the third to context validity, the fourth to scoring validity, the fifth to consequential validity, and the last one to criterion-related validity. Taylor (2011) asserts that at the center of all language testing activities is a triangular relationship between the test taker's cognitive abilities, the task and the context, and the scoring process, which are reflected in cognitive validity, context validity and scoring validity, respectively. She argues:

> By maintaining a strong focus on these three components and by undertaking a careful analysis of their tests in relation to those three dimensions, test providers should be able to provide theoretical, logical and empirical evidence to support validity claims and arguments about the quality and usefulness of their exams. (p. 31)

The focus in the following sections will mainly be on these three components, but first test taker characteristics will be addressed. Consequential validity and criterion-related validity will not be addressed here, since they are beyond the scope of the present study.

2.6.1 Test taker characteristics

Bachman (1990) refers to personal attributes of test takers as test taker characteristics (TTCs). TTCs may be systematic in that they will have an impact upon the test taker's performance under all conditions, even though, as O'Sullivan and Green (2011) suggest, the nature of the effect may differ from one individual to another or depending on the context. These systematic characteristics include "individual characteristics such as cognitive style and knowledge of particular content areas, and group characteristics such as gender, race and ethnic background" (Bachman, 1990, p. 164). Unsystematic or random characteristics, on the other hand, are those that may affect performance on a particular situation, but are not consistent. Since TTCs may affect performance, it is important to collect information on the test taker

population. As O'Sullivan and Green (2011) note, "if we are to create test instruments that reflect the needs and interests of test takers, we need to place them at the centre of our testing circle" (p. 36).

This concern has been raised by a number of scholars. It is argued by Weir (2005) and Taylor (2011) that test tasks should be designed with the purpose of engaging test takers in the cognitive processes that are necessary for communicating effectively in the target language use domain. O'Sullivan and Green (2011) propose that test designers need to pay particular attention to any TTCs that may have an impact upon these cognitive processes. It is commonly pointed out that it should be a priority for test developers and test providers to take measures, as far as possible, to avoid favoring or penalizing any subgroup of test takers (Davies, 2010; Kane, 2010; McNamara & Roever, 2006; Taylor, 2011).

O'Sullivan and Green (2011) cite a number of frameworks that have been proposed that describe TTCs (i.e., Alderson, Clapham and Wall, 1995; Bachman and Palmer, 1996, Brown, 1995). Among the TTCs that have been included in these frameworks are age, gender, first language and educational background, cognitive characteristics, motivation, and test wiseness, to name a few. A more comprehensive model, however, has been offered by O'Sullivan (2000, cited in O'Sullivan & Green, 2011). This model is presented in Table 3. The TTCs in the present study will be investigated in the light of this model.

Table 3. Test Taker Characteristics.

Physical/Physiological	Psychological	Experiential
Age	Personality	Education
Gender	Memory	Examination preparedness
Short-term ailments	Cognitive style	Examination experience
Longer-term disabilities	Affective schemata	Communication experience
	Concentration	TL-country residence
	Motivation	Topic knowledge/
	Emotional state	Knowledge of the world

Source: O'Sullivan and Green, 2011.

2.6.2 Cognitive validity

This section focuses on the cognitive validity of speaking tests drawing upon Field's (2011) adaptation of the cognitive processing model devised by Levelt (1989). Field (2011) provides a number of explanations as to why he uses the Levelt model as the basis for the cognitive validation of L2 speaking tests: it is the most comprehensive account of L1 speech production, as well as the most widely cited, and it is a staged model, which makes it a useful tool in analyzing test specifications in a systematic manner.

Field (2011) notes that cognitive validity of language tests relate to considerations regarding how well the tasks, test content and performance conditions elicit mental (cognitive) processes resembling those which language users employ when undertaking similar tasks in non-test conditions. A second consideration is how finely the relevant processes are reflected in various tasks to cover the intended proficiency levels in terms of the cognitive demands they impose on the test taker. In other words, the tasks are analyzed to see the extent to which they vary and are graded with regards to the cognitive demands they require from the test taker. In that way, a cognitive model of the construct is provided. Field (2011) argues that "this aspect of validity is especially critical in the case of tests that assess general competence in any type of skill across a proficiency continuum" because "the concern of these exams is with the test taker's ability to apply the four language skills at varying levels of proficiency, rather than with the simple measurement of linguistic knowledge" (p. 66).

The focus in studies of cognitive validity is not on the language produced by test takers, but instead on what cognitive processes they go through from the planning of an idea to the articulation of it; therefore, linguistic difficulty is a concern

only in terms of cognitive effort that the retrieval of vocabulary or the use of syntactic structures requires, such as when a comparison is made between familiar vs. unfamiliar vocabulary or simple vs. complex structures (Field, 2011).

2.6.2.1 Field's cognitive processing framework for speaking

Field (2011) builds his framework for a priori cognitive validation of speaking tests on Levelt's (1989) model of speech production. Levelt (1989) views talking as an intentional activity which incorporates a number of processes. These involve conceiving of an intention, accessing the relevant information whose expression will realize this intention, structuring the relevant information, controlling the discourse produced so far both in the whole interaction and by the speaker him/herself, as well as monitoring his or her utterances. As a result, Levelt (1989) asserts that models of speech production need to cover a number of stages. Field (2011, p. 73) presents a condensed explanation for each stage: a conceptual stage (an idea enters the mind of the speaker), a syntactic stage (the appropriate frame for the proposition to be expressed is selected), a lexical stage (the speaker searches his or her lexicon or vocabulary store for expressing the intended meaning and the correct form), a phonological stage (the abstract information formed in the earlier stages takes a speech-like form), a phonetic stage (such features as assimilation are introduced), and an articulatory stage (the message is uttered) . Field (2011) modifies Levelt's model to create a framework to be used for cognitive validation of speaking tests. The stages Field draws upon are conceptualization, grammatical encoding, phonological encoding, phonetic encoding, articulation and self-monitoring. These stages are briefly introduced below.

Conceptualization: Levelt (1989) notes that two types of operations are relevant to conceptualization. These are:

- macro-planning, which is 'the speaker's elaboration of a communicative intention by selecting the information whose expression may realize the communicative goals,
- micro-planning, which is 'planning and informational perspective for an utterance' (Levelt, 1989, p. 5).

Based on these two operations, Field (2011, p. 88) argues that 'conceptualization' can be operationalized in terms of L2 testing criteria under two main headings: provision of ideas and integrating utterances into a discourse framework. With regard to the provision of ideas, Field suggests that test developers can adjust the cognitive demands of a task, first by reducing/increasing the need for conceptualization so that the allocation of working memory to the retrieval of linguistic forms is altered; and secondly, by the support provided to the test taker. As for the integration of utterances into a discourse framework, Field (2011) lists the factors that are suggested by Levelt (1989) to influence both macro- and micro-planning: awareness of the ongoing topic, thematization of new information, recognition of information shared and not shared with the listener, accommodation to the point-of-view and form of the words of the interlocutor. In this regard, the consideration with respect to cognitive validity is "the extent to which test takers are assessed on their ability to relate utterances to the wider discourse" (Field, 2011, p. 88).

Grammatical encoding: Field (2011) suggests that the formulation phase as is defined by Levelt (1989) falls into two parts: a syntactic stage (the appropriate frame for the proposition to be expressed is selected), and a lexical stage (the speaker searches his or her lexicon or vocabulary store for expressing the intended meaning and the correct form). He writes, "Syntactic complexity is clearly a factor in the cognitive difficulty of producing an utterance" (p. 79). However, Field (2011) views

this complexity not only as deriving from linguistic knowledge, but as being influenced by the demands of forming the structure and retaining it in the mind, which clearly relates to cognitive processing. He points out that it is not the grammatical structure, but instead the language functions that test takers are expected to perform that generally determines the linguistic content of speaking tests. With regard to cognitive validity, the issue is not, therefore, the complexity of language in itself but "the *mapping* from the function that the test taker wishes to perform to the pattern that best expresses that function" (Field, 2011, p. 79). The cognitive demands of a task, following this line of reasoning, can be adjusted in two ways: reducing/increasing the number of functions expected from the test taker in a single task, and including either more readily accessible functions or those that are likely to be more difficult to access. Field (2011) suggests that familiar, frequent and concrete functions will be mapped more rapidly and reliably by test takers.

Phonological encoding: Field (2011) suggests that the phase when second language speakers retrieve phonological forms from memory to produce a pre-planned utterance is the most crucial phase of production, because, unlike first language speakers for whom the process is rapid and automatic, they will have to pay higher levels of attention and exert more effort, especially at lower levels of proficiency. The discussion of phonological encoding relates in second language testing to the discussion of fluency, since the ease with which speakers access and formulate word forms generally have an impact upon judgments of fluency. In addition to the discussion of 'fluency' under *What is Speaking* (Section 2.2.), I will include here the physical characteristics listed by Field, which he argues "are associated with progress in the acquisition of the L2 speaking skills and which together form possible indicators of fluency" (p. 83):

- use of pre-assembled chunks – leading to syntactic accuracy and native-like rhythmic properties
- length of run
- duration of planning pauses at syntactic boundaries
- frequency of hesitation pauses. (p. 83)

However, it needs to be noted that fluency relates to cognitive processes not only in terms of the retrieval of phonological forms. No matter how fluent they are, all speakers need to, for example, pause to plan what they will say next (conceptualization) or monitor what they have said before (self-monitoring). This planning process, too, makes cognitive demands from the speaker. It is, therefore, important to consider the reasons for pauses and hesitations (Fulcher, 1996). To sum up, phonological encoding can be operationalized through the fluency scale in speaking assessment; or more precisely, through the gradation of the fluency scale based on what contributes to dysfluency at lower levels or to fluency at higher levels.

Phonetic encoding/ Articulation: Levelt (1989) notes that "fluent articulation is probably man's most complex skill. It involves coordinated use of approximately 100 muscles, such that speech sounds are produced at a rate about 15 per second" (p. 413). He also points out that "the interface of phonological encoding and articulation involves a system that can temporarily store a certain amount of phonetic plan" (p. 414). To put it simply, just as syntactic planning imposes cognitive load on the speaker in relation to the working memory, so does phonetic planning. To these, Field (2011) adds the phonological representations of the L1 and the automatic processes that are developed in the L1 in terms of articulation, which may create greater challenges for L2 speakers. The phonetic encoding and articulation phases are related to the considerations of pronunciation in the case of second language speaking tests. The gradation of pronunciation skills is argued by Field (2011) to rely "chiefly upon a gradual reduction of elements which reflect the influence of the

stored values of L1 and the automatised articulatory processes that are associated with them" (p. 96).

Self-monitoring: In the final stage of the Levelt (1989) model, speakers monitor their utterances (e.g. grammaticality) and how they are producing these utterances (e.g. pragmatic appropriacy). When they realize a mistake, for example, whether this is a grammatical or pragmatic mistake or a mistake in word choice, they may make a repair. In self-monitoring, speakers may pay attention to the information they want to express, how they convey this information, social standards such as formality or register, lexical choices, syntactic and morphological structure, phonology, as well as speed, precision, and fluency. In other words, they monitor the processes that have been discussed so far (conceptualization, grammatical encoding, phonological encoding, phonetic encoding, and articulation). Levelt (1989) concludes that it is extremely unlikely that a speaker can monitor all the levels under the time constraints imposed by the speaking situation, based on "data to support the view that (i) much production trouble is not noticed by the speaker, that (ii) monitoring is context-sensitive… and (iii) a speaker's degree of attention for self-generated trouble fluctuates in the course of an utterance" (p. 463). Field (2011) suggests that since second language speakers have to attend to the challenges of assembling speech, which is highly automatic in the L1, they are likely to limit their self- monitoring. Similarly, Skehan (2009) notes, "If performance in each of these areas, complexity, accuracy, and fluency (CAF), requires attention and working memory involvement, then committing attentional resources to one may have a negative impact on others" (p. 2). These tensions between form and fluency is portrayed by Skehan (1998) in his Trade-off Hypothesis, which predicts that committing attention to one area might lower performance in others. Therefore, test takers' allocation of attention on one

feature of their talk may affect others. Although complexity, accuracy, or fluency can be operationalized in a testing situation, it may prove more difficult to identify how test takers distribute their cognitive and attentional resources. In this regard, Field (2011) notes that "like articulation, self-monitoring and self-repair are aspects of speaker performance which are difficult to capture in the form of test specifications" (p. 96).

2.6.2.2 Time pressure and reciprocity conditions

Field (2011) adds two other factors that may add to cognitive task demands based on Bygate (1987). Bygate (1987) identifies two important variables that add to the cognitive demands of a task: time pressure and reciprocity conditions. He notes that while in writing, users of a language generally have the freedom to create drafts, spend time on what they are going to write, and revise and change what they have written, when they speak, on the other hand, they have to work under the constraints of time and speak with a degree of spontaneity. As the Levelt (1989) model points out, planning is a part of the speaking exercise; however, it is considerably limited in comparison to writing, and speakers can probably plan a few phrases and not more. The formation of an idea, constructing the syntactic, lexical and phonological features of the utterance to be expressed, and the articulation of sounds all take place in a very rapid process, and therefore self-monitoring and self-repair too are subject to the same constraints of time.

A further cognitive demand results from the part played by the interlocutor in the construction of interaction. First of all, the interlocutors may indicate agreement and understanding; however, it is as likely that they will disagree and show incomprehension, in which case the speaker generally needs to adapt his or her speech to attain negotiation of meaning (Levelt, 1989). There is, therefore, a need to

attend to the needs of the listener and the feedback provided by the listener. Bygate (1987) clearly points to this, writing:

> If the processing conditions act as a limitation on our capacity for expression, affecting perhaps the size of the units we use, reciprocity is the condition which challenges us to show continual sensitivity and ability to adjust our use of the language. (p. 13)

Drawing upon these discussions by Bygate (1987) and Levelt (1989), Field (2011) asserts that task length, familiarity of the content, or the language functions that require a mapping exercise for the appropriate language to be retrieved are not the only factors that impact upon the cognitive demands of a task; the amount of time provided to the test taker to form and assemble utterances, and the relationship between the examiner and the test takers as well as between two test takers in paired tasks also influence the cognitive difficulty of a task.

In terms of the amount of time provided to test takers, Field (2011) argues that pre-planning can decrease the cognitive demand imposed on the test taker in terms of conceptualization (e.g. forming relevant ideas for the ongoing topic, organizing them into a coherent discourse, marking the relationship between these ideas), grammatical encoding (attending to the syntactic structure of the utterances to be expressed, and selecting from the lexicon appropriate vocabulary, morphologic form, etc.), phonological encoding (as syntactic and lexical planning may help produce more fluent speech), and lastly self-monitoring (as it enables monitoring of the utterances produced in terms of the pre-planned target utterance). Ortega's (1999) study on the effect of pre-planning on language performances has similarly shown that pre-planning and repetition result in increased fluency and complexity. However, research on the effect of planning on test performance has not been conclusive. This will be discussed in detail under *Time constraints* in the *Context validity* section.

In terms of patterns of interaction, the cognitive demands on the test taker increases when the interviewer asks open ended questions and monitors the speech, instead of asking closed questions that require formulaic responses. The former, Field (2011) argues, better reflects the processing associated with real world spoken discourse. Another concern, as was discussed in the chapter on test tasks, is that the interviewer has considerable power over the test taker and this produces asymmetric participation (Bachman, 1988; Ross & Berwick, 1992). Field (2011) notes that this is one of the reasons for the introduction of candidate-candidate interaction tasks in second language speaking assessment.

2.6.3 Context validity

Weir (2005) defines context validity as "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). Below the aspects of context validity for speaking (see Figure 2) as outlined by Weir (2005), and later discussed in depth by Galaczi and ffrench (2011) in relation to Cambridge ESOL examinations will be presented along with the related literature. Figure 2 presents the aspects of context validity proposed by Weir (2005).

CONTEXT VALIDITY

SETTING: TASK
- Response format
- Purpose
- Weighting
- Known criteria
- Order of items
- Time constraints

SETTING: ADMINISTRATION
- Physical conditions
- Uniformity of administration
- Security

DEMANDS: TASK
Linguistic (Input and Output)
- Channel
- Discourse mode
- Length
- Nature of information
- Topic familiarity/content knowledge
- Lexical resources
- Structural resources
- Functional resources

Interlocutor
- Speech rate
- Variety of accent
- Acquaintanceship

Figure 2. Aspects of context validity for speaking. Source: Weir, 2005

2.6.3.1 Task setting

Task setting refers to "the setting under which the test activity takes place" (Weir, 2005, p.56). The features of the task setting include response format, task purpose, weighting, known criteria, order of items, and time constraints.

Response format: Response formats refer to the patterns of interaction, such as interview, test taker-test taker discussion, or individual long turn. Galaczi and ffrench (2011) cite a number of studies (i.e., Taylor, 1999; ffrench, 2003; and Lazaraton and Frantz, 1997) in the fields of language testing and second language acquisition that have investigated the influence of various response formats on the quality and quantity of test taker performance. These studies found that the relative contribution of the test taker to the interaction and the range of speech functions increased in paired task format when compared to the individual face-to-face speaking format. Lazaraton and Frantz (1997, cited in Galaczi & ffrench, 2011) and O'Sullivan, Weir, and Saville (2002) focused on the speech functions produced by test takers in tasks requiring different response formats. Lazaraton and Frantz (1997, cited in Galaczi & ffrench, 2011) found that the range and frequency of functions elicited by individual monologic and discussion tasks were limited in comparison to examiner-test taker and test taker-test taker interactions. O'Sullivan et al. (2002) investigated the language functions that different task types elicit, and how the nature and quality of language produced in pair format differ from that elicited in examiner-test taker interview format. They concluded that while the interview and monologic formats elicited predominantly informational functions, the discussion format elicited a range of informational, interactional, and interaction management functions.

All of these studies point to the need to include in tests of speaking different response formats in order to minimize construct irrelevant variation that might result

from the use of a single response format, since that will help not only in eliciting various language functions from test takers, but also to replicate as far as possible the jointly constructed nature of spoken interaction. Galaczi and ffrench (2011) write:

> The use of a range of response methods will ensure that the test is not biased towards one particular response format and will lead to more complex cognitive and strategic processes for the test takers, which in turn will elicit a richer range of language, and will allow for better inferences to be made about a candidate's proficiency in wider real-life contexts. (p. 116)

Task purpose: Luoma (2004) identifies task purpose as one of the key decisions in task development and writes, "speaking tasks can be seen as activities that involve speakers in using language for the purpose of achieving a particular goal or objective in a particular speaking situation" (p. 31). In this regard, it is important in tasks to make clear what the task asks the test takers to do, in order words, to clearly communicate the task purpose. Just as important is for these task purposes to be appropriate for the respective proficiency levels aimed by the test (Galaczi and ffrench, 2011). Therefore, a test that aims at a range of proficiency levels should include various task purposes. In this regard, a number of categorizations have been outlined, which can inform the selection of task purposes. For example, Bygate (1987) makes a distinction between factually oriented talk, which includes description, narration, instruction, and comparison, and evaluative talk, which includes explanation, justification, prediction, and decision. Another categorization is provided by the CEFR (Council of Europe, 2001). The CEFR categorizes language functions into macro-functions (description, narration, commentary, explanation, and demonstration) and micro-functions, such as suggesting, reporting, or greeting.

Weighting: A certain aspect of the test may be perceived as more important based on a theory of language ability or the needs of a program, so that this part of the test or assessment criteria is weighted differently than the other parts or criteria (Galaczi &

ffrench, 2011). In the speaking section of the Language Ability Assessment System (LAAS) Spanish test, for example, increased score points are used for the Grammar scale because of the need to adequately differentiate between different levels of grammar knowledge, and the instructors' perception of the relative importance of grammar knowledge (Lynch & McNamara, 1998; Sawaki, 2007).

Known criteria: This aspect of the task setting refers to the familiarity of test takers with the criteria used in the assessment of their performances. Galaczi and ffrench (2011) propose that even though there does not seem to be much evidence in the literature, familiarity with the assessment criteria is likely to affect test takers' performance.

Order of tasks: With regards to order of tasks in a language test, Galaczi and ffrench (2011) argue that it is a common practice of exam boards to start tests with easier, more familiar, and less cognitively demanding tasks, while acknowledging the lack of empirical evidence that task order may affect test taker performance.

Time constraints: A series of studies into the effect of planning time (giving time to the test takers before the task so that they plan what they are going to say) on second language performance have put forward that planning time influences performance more in terms of complexity and fluency in comparison to accuracy (Ortega, 1999, Mehnert, 1998; Yuan & Ellis, 2003). Ortega (1999) reported that the mean number of words per utterance and speech rate were higher in the planned narratives than in the unplanned. Planning time did not have significant effects on accuracy in that study. Ortega (1999) noted, "the pattern of inconclusive findings regarding accuracy observed in previous studies were obtained once again in the present study" (p. 132). Mehnert (1998) studied the influence of different amounts of planning time on performance and found that accuracy was enhanced with one-minute planning, but it

did not benefit from longer amounts of time. Complexity, on the other hand, improved when ten minutes of planning time was provided. Yuan and Ellis (2003) investigated the effects of pre-task planning and on-line planning (planning during the task) on complexity, accuracy, and fluency. They noted that pre-task planning resulted in greater complexity, while on-line planning advantaged both complexity and accuracy. Moreover, pre-task planners produced more fluent language than on-line planners.

On the other hand, Mehrang and Rahimpour (2010) found that planning time had no effect on accuracy and fluency, but it advantaged complexity in unstructured tasks. Another investigation into the effect of planning time on fluency, complexity, and accuracy has concluded that planning time does not make any difference for test takers with higher proficiency on easier tasks; however, in more difficult tasks, as the cognitive demands imposed on the test taker increase, some differences are observed (Wigglesworth, 1997). High-proficiency test takers in the Wigglesworth (1997) study were found to benefit from planning time in terms of complexity, fluency, and accuracy on more difficult tasks; while low-proficiency test takers did not appear to benefit from planning time in terms of complexity, although tentative support for increased fluency was noted in one task and accuracy appeared to benefit from planning time only in the case of indefinite article. On the other hand, in a number of other studies, no impact of the planning time on task performance was observed (Elder, Iwashita & McNamara, 2002; Elder & Wigglesworth, 2003; Iwashita, McNamara & Elder, 2001; Wigglesworth & Elder, 2010).

These observations indicate that planning may enhance performance in cognitively demanding tasks, but the relationship between planning time and performance enhancement, as is indicated by the studies reviewed above, is not

linear. Skehan (1998) points to the need to include tasks with planning conditions and tasks which do not allow for planning, so that more representative evidence can be accumulated regarding test takers' ability to function in a broad range of real-life conditions. Similarly, Elder et al. (2002) propose that a test of academic language ability should include planning time for authenticity reasons, and Wigglesworth and Elder (2010) argue, "planning time is worth taking notice of, if only for face validity reasons" (p. 18).

2.6.3.2 Task demands

Task demands relate to the features of the tasks in terms of linguistic input given to the test takers and the linguistic output expected of them. These features include channel of communication, discourse mode, input length, nature of information, topic familiarity/content knowledge, lexical resources, structural resources, and functional resources.

Channel of communication: This feature refers to the medium through which communication takes place, such as face-to-face communication, or communication through telephone/computer or visuals. Galaczi and ffrench (2011) list the typical channels of communication in speaking tasks as:

- aural (input from examiner or from a recorded medium)
- written (a text to be read by the candidate, usually written at a level of language lower than that of the typical candidate at the proficiency level)
- visual (photos, drawings, pictures, etc.)
- graphical (charts, graphs, tables, etc.). (p. 139)

Field (2011) points out that in semi-direct tests, in which the input comes from a recorded medium and the test taker responds to pre-recorded prompts, there are no facial expressions or gestures that could provide feedback to the test taker to indicate comprehension or no opportunities for the test taker to seek support in cases of, for example, misunderstanding or breakdown of communication. In comparison, in

direct tests of speaking, the construct assessed is spoken interaction, and communication is jointly constructed (either by the examiner and the test taker or among the test takers themselves). Thus, test takers have the opportunity to indicate mis- or non-understanding.

Galaczi and ffrench (2011) cite O'Keefe (2006a and 2006b) who looked into how different channels impacted upon candidate performance, by changing input channel from visual to verbal mode. The results indicated that while the amount of hesitation and use of pre-fabricated language increased in the picture prompts, coherence and cohesion, quantity of lexical output, and syntactic complexity decreased. Galaczi and freench (2011) explain this by the cognitive load imposed in the picture prompts since the test takers need to, for example, compare pictures and retrieve related lexical items at the same time.

Discourse mode: Galaczi and ffrench (2011) point to the lack of agreement in the literature on the terminology used to classify various discourse modes, writing "traditionally discourse mode refers to the modes of narration, description, exposition, and argument/persuasion" (p. 143). Discourse modes have been categorized in various, albeit similar, ways: description, instruction, storytelling and opinion expressing/justification (Brown & Yule, 1983); factually oriented talk which include description, narration, instruction, comparison, and evaluative talk which include explanation, justification, prediction, decision (Bygate, 1987); and description, narration, commentary, explanation and demonstration (Council of Europe, 2001). Discourse mode is important to take account of in test situations, since different discourse modes may present different levels of difficulty for test takers (Galaczi & ffrench, 2011).

Input length: This feature refers to the length of the instructions given to the test takers. Galaczi and ffrench (2011) suggest that as the length of task input increases, so does the cognitive demand on test takers.

Nature of information: This aspects of the task relates to the type of information required by the task. It is related to task difficulty in that as tasks move from personal to non-personal or from concrete to abstract, their difficulty is suggested to increase as well (Skehan, 1998).

Topic familiarity/Content knowledge required: With regard to background knowledge, O'Sullivan (2008) argues:

> Background knowledge has not been considered a major factor in the testing of speaking. This is because of the task types typically used, where knowledge is presented in the task input in order to ensure that all language samples elicited will be comparable. While there are some examples of so-called 'free conversation' type oral examinations, the general consensus nowadays is that the inherent inequality of the test event makes true *conversation* (where all interlocutors have equal control over the discourse) impossible. (p. 8)

On the other hand, it has been suggested various times in the literature that in tests of general proficiency, the content should not be biased towards a certain sub-population of test takers (Hughes, 2003; Weir, 1993). Alderson (2000, cited in Galaczi & ffrench, 2011) notes, in relation to reading assessment, that "every attempt should be made to allow background information to facilitate performance rather than allowing its absence to inhibit performance," (p. 29) which equally well applies to the assessment of speaking skills. However, different findings have been noted in the literature with regard to the impact of topic familiarity on test performance. Douglas and Selinker (1992) noted that it should not be readily assumed that test scores will increase when candidates are allowed to talk about matters from their own field. Lumley and O'Sullivan (2005) deliberately manipulated a number of tasks to investigate whether task topics would result in bias for either gender and

concluded that most of the bias terms were small, and when bias was detected, this only appeared to affect a single scoring category. For example, in the task on basketball small effect size was found, while topics about soccer had no adverse effect on female test takers. They concluded that "the bias is not always predictable: instances are observed here where there was unexpected bias, and others where it failed to be seen where it might be anticipated" (p. 432).

Even though research has so far not been conclusive with regards to the effect of background knowledge on test performance and test scores, it is essential that every attempt be made to avoid bias in task topics against any group of test takers. In this regard, Galaczi and ffrench (2011) list a number of principles taken from Cambridge ESOL Item Writer Guidelines for PET – CPE. The guidelines advise that topics are:

- appropriate for the levels of candidature from all cultures, experiences and age groups
- within the age group and experience of the candidature
- free of any cultural bias (e.g. urban/rural, boy/girl, cultural, etc.)
- not potentially distressing or offensive, such as alcohol, drugs, war, death, politics, religious beliefs, racism, sexism, etc.
- likely to appeal to a broad base of candidates and will not 'date' too quickly
- able to provide a range of options in terms of subtopics and functions
- at an appropriate cognitive level and are likely to produce answers of the appropriate level and of the required length for the particular candidature
- familiar, yet not too familiar, i.e. the candidates should have an existing schema (organized mental framework) for the topic
- not specialized or technical, so that the material does not favor candidates with specialized knowledge of a particular subject. (p. 150)

Lexical resources, structural resources, and functional resources: It has been noted in the literature that as the proficiency level increases, so does the number of word types and tokens (Iwashita, Brown, McNamara, & O'Hagan, 2008). However, Davis (2009) notes that "higher word count per se does not appear to generate higher scores" (p. 387). In terms of frequency of the vocabulary items used by candidates too, different results have been reported. While Laufer and Nation (1999) suggest

that the use of lower frequency words increases with proficiency level, Schmitt (2009, cited in Galaczi & ffrench, 2011) found similar frequency profiles across different proficiency levels.

Galaczi and ffrench (2011) discuss lexical, structural, and functional resources based on the CEFR (Council of Europe, 2001). In terms of lexical resources, The CEFR indicates that learners at A2 can control a narrow repertoire dealing with particular concrete situations; by B1 they can express themselves on topics about everyday life such as hobbies and interests; B2 leaners have a command of a good range of vocabulary about their field and most general topics; and at C levels, they have broad lexical repertoire including idiomatic expressions and colloquialisms. In terms of structural resources, learners at A1 and A2 levels have limited control of structural resources; B1 is characterized by frequently used structures; by B2 a high degree of grammatical control is attained; and at C levels, learners can use a wide range of resources with a consistent degree of control.

In terms of structural resources, the CEFR indicates that A1 learners have limited control of a few simple grammatical structures, and A2 learners still systematically make basic mistakes. At B1 level, they make reasonably accurate use of frequently used structures, and by B1 there is a high degree of grammatical control. At C levels, learners consistently maintain a high degree of grammatical accuracy and they make use of a broader range of structural resources.

Functional resources, as has been noted under the section titled *Discourse Mode*, are discussed in the CEFR in relation to macro- and micro-functions.

2.6.3.3 Interlocutor effects

Interlocutors may function as another source of score variance in direct tests of oral language ability. In interview tasks, both Brown (2003) and Nakatsuhara (2008)

observed differences between interviewers in terms of their questioning, extending topics, and providing feedback. Moreover, these differences were reflected in the scores assigned to the test takers. An interesting observation made by Nakatsuhara (2008) was that the more trained interviewer led the candidate to appear to perform less well.

Another line of research investigated the effect of pairings in tasks that require test taker-test taker interaction. Davis (2009) reported that the proficiency level of one's partner did not produce significant differences in terms of the amount of language produced or the scores received. Norton (2005), on the other hand, found that linguistic performance and the amount of talk produced are potentially affected by pairing, if test takers with different levels of proficiency are paired or if test takers know each other. However, she wrote, "Given the numerous and interrelated variables on the assessment process, it is impossible to show a definitive correlation between pairing and the final marks awarded" (p. 287).

2.6.3.4 Administration setting

This feature relates to the setting under which test administration takes place. Galaczi and ffrench (2011) note that "administrative and environmental conditions can also impact on both Cognitive validity (i.e. mental processing) and Scoring validity (i.e. quality of rating), especially where direct speaking assessment is concerned" (p. 336). They point particularly to the importance of equivalence in test administrations. They argue that although administrative setting could be located as an aspect of context validity (the setting for the task), this would be an oversimplification. This aspect is viewed as being related to other aspects such as timetabling or supervision of candidates, in addition to task setting variables.

So far, the aspects of tests that relate to a priori validation have been reviewed. In the next section, a number of issues that arise following test administration will be covered.

2.6.4 Scoring validity

Taylor and Galaczi (2011) note that the phrase 'scoring validity' was first adopted by Weir (2005) to cover all aspects of the assessment process which may influence the consistency and dependability of scores. In other words, it was coined as an umbrella term for all aspects of reliability. Shaw and Weir (2007) write the following:

> [Scoring validity] accounts for the extent to which test scores are based upon appropriate criteria, exhibit consequential agreement in marking, are as free as possible from measurement error, stable over time, consistent in terms of content sampling and engender confidence as reliable decision-making indicators. (p. 143)

Then, they list a number of factors that may threaten the scoring validity of an assessment such as imprecise scaling descriptors, inconsistent raters or unsuitable rating procedures, inadequate training or unsystematic grading, to name a few. To these Taylor and Galaczi (2011) add multiple other factors drawing upon Galaczi and ffrench (2011), Field (2011), and Shaw and Weir (2007). These include a limited range of task and interaction types, allocation of different amounts of time and weigh to different components of the test, variation in terms of channels of communication, variables that stem from the interlocutor, imprecise or ambiguous rubrics, or inconsistency across different administrations of the test. They argue, "Given the wide range of factors which can impact on the scoring validity of a speaking test, the notion that speaking test reliability can be adequately captured and expressed in the form of a single statistic, typically an inter-rater correlation coefficient, is not appropriate," (p. 172) suggesting that in addition to rater variables which are expressed in an inter-rater correlation coefficient, other variables that can influence

scoring validity need to be addressed. Based on Weir's (2005) socio-cognitive validation framework, they list the following parameters as relating to scoring validity:

- Assessment criteria/ rating scale
- Rating process
- Rating conditions
- Rater characteristics
- Rater training
- Post-exam adjustment
- Grading and awarding

2.6.4.1 Assessment criteria and rating scales

Various researchers have commented on the importance of rating scales in the assessment of writing and speaking abilities (Alderson, Clapham & Wall 1995; Bachman & Palmer, 1996; McNamara, 1996). However, before discussing the uses of rating scales, I will provide a definition of 'rating scale' and discuss different types of them. A rating scale can be defined as:

> A scale for the description of language proficiency consisting of a series of constructed levels against which a language learner's performance is judged. Like a test, a proficiency (rating) scale provides an operational definition of a linguistic construct such as proficiency. Typically such scales range from zero mastery through to an end-point representing the well-educated native speaker. The levels or bands are commonly characterized in terms of what subjects can do with the language (tasks and functions which can be performed) and their mastery of linguistic features (such as vocabulary, syntax, fluency and cohesion)… Scales are descriptions of group of typically occurring behaviors; they are not in themselves test instruments and need to be used in conjunction with tests appropriate to the population and test purpose. Raters of judges are normally trained in the use of proficiency scales so as to ensure the measure's reliability. (Davies, Brown, Elder, Hill, Lumley, & McNamara, 1999, p. 153-154)

The development of rating scales, however, is not as simple as it may seem. North (2000, cited in Luoma, 2004) points to this issue, noting that rating scale developers

are in fact “trying to describe complex phenomena in a small number of words on the basis of incomplete theory” (p. 13). Luoma (2004) emphasizes the inadequacy of solid evidence about language learning and the challenge of summarizing descriptors into brief statements, but adds that scales function as a part of construct definition and are developed to differentiate between good and weak performances, thus rendering them an essential part of the assessment process. Rating scales are important also because “the band descriptor is a major part of the ‘meaning’ of the score, and delimits the type of inferences that can be made from the test score” (Fulcher, 2003, p. 113).

Rating scales have been divided into two main types. The first is behavioral-oriented and hence defines language ability in relation to real life performance, focusing on language use in specific contexts. Bachman (1990) refers to this approach as the “real-world” approach (p. 344). This approach has been subject to criticism on the basis that behavioral scales limit the generalizability of score interpretation in that it is not possible to generalize from test scores to speaking situations that are not captured in test tasks, and that they are inadequate in providing information about various components of language ability (Bachman & Savignon, 1986; Bahman & Palmer, 1996). The second type defines language ability independently of context of language use, and is based on a theoretical model of the construct that is being assessed. This approach relates more to constructs than test tasks. In other words, it is construct-oriented. It is suggested that generalization beyond the tasks is possible when construct-oriented rating scales are used (Bachman, 1990; Bachman and Palmer, 1996).

Another distinction between rating scales is based on their uses. Alderson (1991) suggests that rating scales can be user-oriented, assessor-oriented, or

constructor-oriented, based on who the scale aims to inform and who will use the scale. He notes that while user-oriented scales have a reporting function, assessor-oriented scales guide the rating process, and constructor-oriented scales guide the construction of tests at appropriate levels.

The last distinction between scales relates to the scoring approach taken up by a rating scale: holistic vs. analytic. When holistic scales are used, the rater makes an impressionistic assessment of a test taker's ability using a single rating scale (Davies et al., 1999). The rater may pay attention to different features of the quality of spoken language performance in the case of speaking assessment, or to overall impression only. The advantages and disadvantages of this approach have been noted by several researchers (Barkaoui, 2007; Fulcher, 2003; Luoma, 2004; Weir, 1993). Luoma (2004) writes that holistic scoring is practical, and can be undertaken relatively quickly, pointing out at the same time that "they are not practical for diagnosing strengths and weaknesses in individual learners' performances" (p. 62). Galaczi and ffrench (2011) write that holistic descriptors can provide a summary of skill levels that can be intuitively accessed and used without requiring any specialist knowledge of linguistics or language assessment. However, raters may focus on different aspects of test taker performance in arriving at judgments using self-generated criteria (Weir, 1993; Barkaoui, 2007). Moreover, Weir (1993) points out that holistic scales are generally developed intuitively, and therefore, they lack empirical foundation.

When analytic scales are used, on the other hand, raters are required to attend to a number of criteria, and provide a score for each criterion for individual test takers. Luoma (2004) writes that analytic scales have advantages such as providing detailed guidance to the raters and information on specific strengths and weaknesses of test takers, but warns that when raters are required to make multiple judgments,

this means increased cognitive load. The CEFR (Council of Europe, 2001) suggests that seven categories is the psychological upper limit for raters. Another advantage of analytic scales is that since they draw raters' attention to specific features, they can contribute to rater agreement and rater reliability unless the criteria are inexplicit or vague (Weir, 1993). Hughes (2003) expresses similar advantages of analytic scoring:

> It disposes of the problem of uneven development of subskills of individuals. Secondly, scorers are compelled to consider aspects of performance which they might otherwise ignore. And thirdly, the very fact that the scorer has to give a number of scores will tend to make the scoring more reliable. (p. 102)

Galaczi and ffrench (2011) note, however, a number of disadvantages of analytic scales. In the first place, the rating criteria may not be functioning independently or raters may not distinguish between them for various reasons. In addition, analytic scoring is time consuming and may not be practical in certain assessment contexts, such as when the rater is at the same time the interlocutor in a test.

The analytical criteria included in speaking tests can be exemplified as follows: fluency and coherence, lexical resource, grammatical range and accuracy, pronunciation (International English Language Testing System, 2009); grammatical resource, lexical resource, interactive communication, discourse management, pronunciation (University of Cambridge ESOL Examinations, 2015); delivery, language use, topic development (Educational Testing Service, 2004); range, accuracy, fluency, interaction, coherence (Council of Europe, 2001).

Based on the distinctions made between categorizations of rating scales presented above, Fulcher (2003) developed a framework for describing rating scales. According to this framework, rating scales are of three types in terms of orientation: user-, assessor- and constructor-oriented scales. They can be based on an analytic approach or holistic approach to scoring, and holistic approach includes holistic

scoring, primary-trait scoring, and multiple-trait scoring. Lastly, they can be real world-focused or construct-focused.

There are a number of concerns that need to be addressed in rating scale development. These include decisions regarding the number of levels and the number of criteria to be used in the scale and descriptions of score meanings (Luoma, 2004). Luoma (2004) points out that even though more levels can provide more detailed information, care needs to be taken to ensure that raters are able to distinguish the levels consistently. Moreover, the criteria should be conceptually independent. She suggests that five to six criteria may be close to the maximum. With regards to the number of levels, Isaacs and Thomson (2013) noted that while 5-point scales were too constraining for some raters, when 9-point scales were used, raters were unable to differentiate between the levels. In addition to having a justifiable number of levels and criteria, the scale descriptors should be concrete, practical, and easy for raters to memorize (Luoma, 2004). Another concern is raised by Fulcher (1996), who criticizes intuitive construction of scales, and instead favors scale development based on empirical performance data.

2.6.4.2 Rating process and rating conditions

The rating process and rating conditions parameters relate to the decision making processes and temporal, spatial, or psychological conditions under which raters need to operate while scoring performances (Taylor & Galaczi, 2011). Taylor and Galaczi (2011) argue that these parameters naturally differ and affect the rating, for example, when raters have to make real-time judgments as compared to rating video-taped performances, in which case they would have the opportunity to replay the video if they needed to. Similarly, they point out that a rater who participates as an

interlocutor in the testing process would have more cognitive load in comparison to one who does not.

2.6.4.3 Rater characteristics and rater training

Research has shown extensive variability in test scores that are attributable to rater effects. Rater effects such as severity or leniency, halo, central tendency, randomness, and bias are viewed as sources of systematic error variance (Eckes, 2005; Hoyt, 2000; Myford & Wolfe, 2003, 2004). Substantial variation in rater severity has been noted in various research studies (Chalhoub-Daville & Wigglesworth, 2005; McNamara & Adams, 1991; Weigle, 1998) and this variation has been observed even after extensive rater training (Eckes, 2009; Lumley & McNamara, 1995; Weigle, 1998), although it should be noted that other research showed some positive effects of rater training (Elder, Knoch, Barkhuizen & von Randow, 2005; Johnson & Lim, 2009; Kang, 2008; Weigle, 1994; Xi & Mollaun, 2009, 2011). Some of the studies that have addressed rater characteristics as a source of score variation are reviewed below.

One line of research has investigated the effects of raters' native language on their ratings of test taker performances. Non-native speakers have been observed to be more severe in their assessment of writing performances than native speakers (Shi, 2001). Other research, similarly, pointed to differences between raters from different national and linguistic backgrounds (for example, native speakers of English from different countries of residence, or rater familiarity with the test takers' first language) in terms of their evaluation of learners' oral performances (Carey, Mannell, & Dunn, 2011; Chalhoub-Deville & Wigglesworth, 2005; Kang, 2008). However, a number of studies found the opposite. Kim (2009) and Zhang and Elder (2010) found only marginal differences in the consistency or severity of native and

non-native raters although both noted that native speaker raters were more detailed in their comments. Brown (1995) found little evidence that native speakers would be more suitable than nonnative speakers, and it has been suggested that raters from different languages can be trained to be as effective as native speaker raters (Johnson and Lim, 2009; Xi & Mollaun, 2009, 2011). In terms of rater's familiarity with test takers' first language, Winke, Gass and Myford (2013) found that L1 familiarity contributed to rater leniency, but the effect sizes were small. To sum up, as Zhang and Elder (2010) note, "no consensus regarding the effect of language background on rating behavior has yet been reached" (p. 4).

However, as was pointed out above, rater characteristics have been consistently found to affect test scores. Significant effects of bias towards specific test takers (Kondo-Brown, 2002; Lynch & McNamara, 1998), specific test tasks (Lynch & McNamara, 1998; Wigglesworth, 1993) and specific rating criteria (Eckes, 2009; Wigglesworth, 1993) have been observed. One obvious implication of this is the need to address this variability by rating training. As noted earlier, however, even though positive effects of rater training have been extensively reported in the literature, it has also been shown that rater training does not completely eliminate rater variation. McNamara (1996) suggests that rater variability is a "fact of life", adding that it may be unachievable to eradicate differences between raters (p. 127). Another way to deal with rater variability, then, is to use statistical tools such as Many-Facet Rasch Measurement to adjust score variance produced by assessment conditions other than the test taker (see Eckes, 2011 for an introduction to Many-Facet Rasch Measurement) or to improve the assessment design by investigating the ideal number of raters for score reliability. The latter can be achieved by using

generalizability and decision analyses. The following section will provide a brief introduction to these analyses, since these analyses were utilized in the current study.

2.7 Generalizability theory

In measuring proficiency in a second language and linking observations of test performances to interpretations about a test taker's ability to use a language in a particular context or for a particular purpose, language testers aim to be able to make generalizations beyond the test. These inferences and generalizations are closely connected to issues of validity and dependability (reliability)[1] of test scores (Deville & Chalhoub-Deville, 2006). Chalhoub-Deville (2006) writes, "dependability or reliability, in a broad sense, refers to the consistencies of data, scores, or observations obtained using elicitation instruments" (p. 2) and lists a number of sources of error that can limit the degree of reliability. These include the elicitation method, the number of elicitations, and the influence of the interlocutor or observer involved in the elicitation.

Brennan (2000) argues that historically the approach to issues of reliability in psychology and education is predominantly based on the use of classical test theory (CTT). CTT partitions observed-score variance into systematic (also called true score variance) and random variance (also called error variance) (Feldt & Brennan, 1989, cited in Brennan, 2000). In other words, a classical reliability coefficient generally implies a single undifferentiated source of measurement error. Therefore, when the assessment involves a single measurement facet, CTT provides sufficient information regarding the generalizability of test scores (Lee, 2006). However, as Shavelson, Webb and Rowley (1989) point out, the usefulness of CTT "depends on the researcher's ability to estimate true score and error variances from data. With

[1] The terms *reliability* and *dependability* will be used synonymously in the following sections.

practical application of CT, we find that error variance is not a monolithic construct; error arises from multiple sources"[2] (p. 922). Even though it is the most widely used indicator of reliability, reliability coefficient may not always provide the most appropriate estimate, since it may under- or over-estimate reliability depending on how an assessment system is constructed (Deville & Chalhoub-Deville, 2006).

The assessment context introduces a range of factors other than the object of measurement (i.e., test takers) which may have an impact upon the performance of a test taker in second language performance tests (Lynch & McNamara, 1998). Speaking, similarly, involves more than one random facet (Lee, 2006). Research has particularly investigated variability in assessment tasks and rater judgements as random facets that cause error variance in performance testing (e.g. Chalhoub-Daville & Wigglesworth, 2005; Lynch & McNamara, 1998; Wigglesworth, 1993). Such an assessment context as the one in speaking and writing tests, which include various sources of error, requires a multifaceted analysis. In many studies that have examined multiple sources of variability, generalizability theory (G theory) has been applied (Brown, 1999; Brown & Ahn, 2011; Lee & Kantor, 2005). A number of studies focused on applications of G theory to the assessment of speaking (Bachman, Lynch & Mason, 1995; Lee, 2005, 2006; Lynch & McNamara, 1998; Sawaki, 2007). Huang (2012) states that G theory has also been used as a method of examining the construct validity of language tests.

Brennan (2000) writes that "generalizability (G) theory liberalizes CTT by providing models and methods that allow an investigator to disentangle multiple sources of error that contribute to E"[3] (p. 339). Shavelson, Webb, and Rowley (1989) list a number of ways in which G-theory extends the framework of CTT:

[2] Shavelson, Webb and Rowley (1989) uses CT as the abbreviation of Classical Test Theory.
[3] E is the undifferentiated random error term.

> First, it recognizes multiple sources of measurement error, estimates each source separately, and provides a mechanism for optimizing the reliability. Second, although GT provides a reliability coefficient, called a "generalizability coefficient," the theory self-consciously focuses on variance components that index the magnitude of each source of error. Third, GT distinguishes between *relative decisions*, where interest attaches to the dependability of the differences among individuals, and *absolute decisions*, where scores are themselves interpretable without reference to others' performance. Fourth, GT distinguishes between *generalizability* (G) and *decision* (D) studies. G studies estimate the magnitude of as many potential sources of measurement error as possible. D studies use information from a G study to design a measurement that minimizes error for a *particular* purpose. (Italics original, p. 923)

Briefly, "G-theory uses the analysis of variance (ANOVA) to partition the variation in scores into different sources and their interaction terms" (Huang, 2012, p. 126). These sources of variance, also called variance components (VCs), are then used to estimate the impact of various changes in assessment design (generally different numbers of raters and tasks in the case of speaking assessment) on the generalizability or dependability (which is analogous to reliability in CTT) of the scores (Brown, 1999; Brown & Kondo-Brown, 2012) through a decision study (D study). So, as Huang (2012) notes, a D study uses the same data as the G study and provides estimations about the relative effects of specified numbers of conditions for each VC (For example, for using one task and two raters, two tasks and one rater, two tasks and two raters, etc.).

CHAPTER 3

METHODOLOGY

3.1 Introduction

In this chapter, the methods and procedures of the present study will be described. The chapter will provide information about the aim of the study, research questions, participants, instruments, data collection procedure, scoring procedure, and analyses of data.

3.2 The aim of the study and research questions

This study aims to investigate and describe the initial steps in the development of a test of speaking for learners of Turkish as a second language, drawing upon the theoretical framework for validating language examinations outlined by Weir (2005). The study deals with aspects of a priori validation of test tasks in terms of cognitive and context validity, as well as aspects of a posteriori validation, with the application of classical test theory and generalizability analyses, using the latter as a basis for investigating the scoring validity of the test. With this aim, the study addresses the following research questions:

1. What are the cognitive demands required by the tasks from the test takers?
2. What are the contextual characteristics of the tasks in terms of performance conditions and linguistic demands?
3. How reliable are the rating process and the scores assigned to the test takers?
4. How dependable are the scores assigned to the test takers?
    4.1. To what extent are the analytic scores dependable?
    4.2. How many tasks and raters would produce relatively more dependable scores?
5. To what extent are the test takers' scores related to the scoring dimensions?

With respect to the first question, the cognitive processing framework for speaking that was developed by Field (2011) based on the Levelt (1989) model of speech production was adopted. An important consideration in this investigation was whether the cognitive demands imposed upon the test takers varied across the tasks, since they aim to cover a range of proficiency levels.

The second question was examined against the aspects of context validity for speaking outlined by Weir (2005). Weir's model of context validity involves three components: task setting, administration setting, and task demands including linguistic and interlocutor factors. The administration and interlocutor variables were not investigated in the current study. The tasks were examined with regard to task setting (performance conditions) as well as linguistic demands of the tasks.

For the examination of the third research question, the aspects of the testing process that relate to the raters and the rating process were analyzed. The raters, rating procedures, and rating conditions are described for standardization of the rating process. To examine the reliability of the scores assigned by the four raters, the consistency among the rater scores were examined.

Regarding the fourth question, generalizability analyses were conducted to assess the dependability of the analytic scores. In addition, decision studies allowed to determine the number of tasks and raters that would produce dependable scores.

The fifth question examined another aspect of scoring validity, that is the consistency of scores across the scoring dimensions (Fluency, Grammatical Range and Accuracy, Lexical Resource, Coherence/Interaction). It was assumed that the relationship among the scores obtained from different dimensions would provide further evidence for the dependability of scores.

3.3 Participants

The participants involved in the present study were 24 learners of Turkish as a foreign language, who were enrolled in Turkish as a foreign language classes at Boğaziçi University at the time of the data collection. They were at Boğaziçi University for either one or two semesters through international student exchange programs. Because two of them did not fill in the Learner Profile Form (see Appendix D), the information provided here will be about 22 of the participants. 16 of the participants were female, and six were male. Their age ranged from 20 to 37, with a mean of 21. Most of them had started learning Turkish for over two years before the data collection, and only five of them had been learning Turkish for less than two years. However, their length of residence in Turkey was relatively short, with most of them living in Turkey for about three months at the time of the data collection. Their average length of residence in Turkey was 15 months; when a test taker who had been living in Turkey for over 10 years was excluded from the analysis, the average length of residence decreased to six months. Their average length of learning Turkish was 67 months; however, when the four participants who had been learning Turkish since childhood were excluded, it decreased to 23 months. The participants formed a heterogeneous group in terms of country of birth, native language, and language of education; so, they had different backgrounds in these regards. Table 4 presents the numbers of participants coming from each country.

The Turkish for Foreigners (TKF) classes that the participants were enrolled in were offered by the Turkish Language and Literature Department of Boğaziçi University. There are different classes for students from different levels of proficiency. TKF 111 and TKF 112 are offered to students with lower levels of proficiency, TKF 211 is for intermediate level, while TKF 315 and 317 are for

advanced students. However, the assignment of students into different classes is not carried out through standardized assessment; instead, the decision is generally made by the course instructors based on their impressionistic judgments of student proficiency (which are based on either face-to-face conversations or student e-mails) or the students themselves choose the classes based on their own perception of their proficiency levels. Therefore, there was no objective evidence as to the proficiency levels of the students before the data collection. The participants of the present study were from TKF 211 (5 students) and TKF 315 (17 students) classes. Participation in the study was on a voluntary basis, and the students from lower levels did not volunteer to take the test, probably because they were not confident they would be able to handle the test tasks.

Table 4. Number of Participants Coming from Each Country.

Nationality	Frequency	Percentage
Germany	6	27,2
Japan	5	22,7
USA	2	9,0
Greece	2	9,0
Jordan	1	4,5
Netherlands	1	4,5
Sweden	1	4,5
Norway	1	4,5
Austria	1	4,5
Italy	1	4,5
Kosovo	1	4,5

The participants were asked to self-evaluate their proficiency through the CEFR Self-Assessment Grid ‘speaking’ and ‘interaction’ sections (Council of Europe, 2001, p. 26-27). For ‘speaking’, TKF 211 group’s (n=5) average reported proficiency level was 2.2 (2 corresponds to A2 on the CEFR scale); and TKF 315 group’s (n=17) average reported proficiency level was 3.5 (4 corresponds to B2 on the CEFR scale). However, as was indicated, there were no tests to assess students’ true level of

proficiency to assign them to different classes and the reports showed there were students in TKF 315 classes who self-assessed themselves on A1 and A2 levels as well as those who reported C2 level of proficiency. For ‘interaction’, TKF 211 group’s average reported proficiency level was 2.2; for TKF 315 group, it was 3.41.

3.4 Instruments

3.4.1 Speaking tasks

The speaking tasks used in the present study were developed by the researcher with guidance from the course instructors and experts. Since the test was intended to be used as a general proficiency test, the test tasks were informed by CEFR ‘can do’ statements from A2 to C2 levels. The test development process included various stages such as selection of task types, writing of task items, consulting with experts, and pilot tests. The current study is based on the second pilot test. The first pilot test will not be discussed here, but references will be made to it where relevant. It should suffice to note here that the tasks used in the present study (see Appendix A) were the revised forms of the initial tasks (with the addition of one new task) following student feedback, expert feedback, and analysis of the first pilot test results.

3.4.2 Rating scale

The rating scale for the assessment of test taker performances in the present study was adapted from the CEFR scale descriptors. It is an analytic scale with five dimensions: Fluency (F), Grammatical Range and Accuracy (A), Lexical Resource (L), Coherence/ Interaction (C/I). The scale was developed through several revisions iteratively. An initial form was developed through the examination of test taker performances. The purpose was to base the scale descriptors on empirical evidence from the performances of test takers with different levels of proficiency. However, based on expert feedback, this initial form of the rating scale was abandoned since

the sample in the present study would not present enough evidence to describe specific proficiency levels. There were only a few students from each level of proficiency and their performance could not be claimed to be representative of a particular proficiency level. For this reason, the latest form of the rating scale was adapted from the CEFR level descriptors. The two forms of the rating scale and an account of the scale adaptation process are provided in Appendix C.

3.4.3 Learner profile form

A Learner Profile Form was developed to collect information from the participants about demographics such as age, gender, country of birth, mother tongue, language of education, duration of learning, and exposure to Turkish or length of residence in Turkey. The students were also asked to self-assess their current level of proficiency in Turkish based on the CEFR Self-Assessment Grid. The information collected through the Learner Profile Form informed the task development process. The form is presented in Appendix D.

3.4.4 Task evaluation form

A Task Evaluation Form was developed to ask for feedback from the participants about the test tasks. The form investigated the participants' opinion on task difficulty, the clarity of the instructions and oral/written prompts, the sufficiency of the time allocated to the tasks, the efficiency of the picture prompts, and the relevance of the tasks to the participants' real life communication and the TKF course contents. The form is presented in Appendix E.

3.5 Data collection procedures

Four of the six tasks were individual tasks where the examiner gave the instructions and asked the questions, and there were two other tasks which required interaction of two participants with each other. Table 5 provides the time allotted to each task. The

amount of time to be allotted to the tasks was decided upon based on average amount of time that the tasks took to complete in the first pilot examination and test taker feedback. There was one new task, Task 6, in the second pilot examination. The amount of time to be allocated to this task was decided by the researcher.

Table 5. Duration of the Speaking Tasks.

Task	Intended level	Time given for the task
1	A2 to C2	7 min.
2	B1	4 min.
3	B2 and C1	5 min.
4	A2 and B1	3 min.
5	B2 to C2	4 min (1 min. for preparation; 3 min. for speaking)
6	A2 and B1	5 min.

Since the students had a tight schedule and since participation was voluntary, the six tasks were divided into two forms, Form 1 (Task 1, 2, 3) and Form 2 (Task 4, 5, 6). The forms and the tasks are presented in Appendix A. Both forms included two individual tasks and one paired task. 12 of the participants took one form of the test, and 12 others took the other form. The allocation of the tasks to the two forms was random. The participants were informed about the test and a short description of each task was sent to the participants via e-mail. Three examiners were recruited for the administration of the tasks. One was the researcher himself, and the other two were colleagues of the researcher, both of whom worked as research assistants at the Department of Foreign Language Education of Boğaziçi University. The examiners were trained about how to implement each task in order to ensure standardization. All of the participants were audio-recorded to be rated afterwards. The participants were expected to fill in the Task Evaluation Form after they completed each task. The tasks were administered over a period of three weeks since the administrations had to be scheduled to time slots when the participants were free. Before the administrations, the researcher visited the TKF classes to distribute a time schedule

to the participants, and they wrote their names and contact information to slots when they could take the test.

3.6 Scoring procedure

For the inter-rater reliability, generalizability, and decision analyses, four raters scored the performance of each candidate on each of the four dimensions included in the rating scale on each of the three tasks (12 test takers took the three tasks in Form 1 and 12 others took the other three tasks in Form 2).

The raters were native speakers of Turkish. They were senior undergraduate students at Foreign Language Education Department, Boğaziçi University. They did not have experience in scoring L2 Turkish performances. An intensive rating training session on the rating scale and scoring procedures took place. Student performances from the first and second pilot tests were used to exemplify performances at different levels for each dimension in the rating scale. After the ratings, the scores were compared by the researcher and when discrepancies were observed, the raters were asked to rate for a second time these particular performances on which there were disagreements. The raters were free to change the scores they assigned in the first rating or give the same score again. In a number of cases, the raters decided to keep the first scores, while in others they assigned a different score. The goal in rater training and monitoring was to minimize differences in ratings.

3.7 Data analysis

For the first research question, the test tasks were analyzed in terms of the cognitive demands they impose upon the test takers, based on the cognitive processing model developed by Field (2011). The model provides a five-part framework for examining cognitive validity: conceptualization, grammatical encoding, phonological encoding, phonetic encoding and articulation, and self-monitoring. Based on Bygate (1987),

Field (2011) proposes two other cognitive factors in addition to those drawn from the Levelt model – time pressure and reciprocity conditions. These were also considered in relation to the tasks investigated under the present study.

Conceptualization involves the complexity of ideas required by and the support provided in each task as well as how the tasks take account of the ability to relate utterances to a wider discourse representation. Specifically it involves:

a) *Provision of ideas* – the complexity of the ideas which test takers have to express and the extent to which the ideas are supplied to them
b) *Integrating utterances into a discourse framework* – the extent to which test takers are assessed on their ability to relate utterances to the wider discourse (including their awareness of information shared with the interlocutor). (Field, 2011, p. 88)

Grammatical encoding refers to the mapping from the function to be expressed to the pattern that best expresses this function. Field (2011) discusses two ways in which the cognitive load imposed upon the test taker can be decreased while performing this mapping exercise. “One lies in restricting the number of functions that a test taker is expected to perform (particularly in relation to a single task)… The other takes account of how accessible a given function is likely to be” (p. 90-91). In line with this, the functions required in each task were listed, and the restriction of functions within the tasks was discussed.

Phonological encoding is examined in the current study in relation to the Fluency rubrics of the rating scale. This section considers how the gradation of the scale takes account of the factors that may be responsible for dysfluency at lower and fluency at higher proficiency levels.

Phonetic encoding and articulation are operationalized by Field (2011) in relation to pronunciation and intelligibility. Pronunciation does not feature among the assessment dimensions in the current test. The reasons for this are discussed.

In terms of self-monitoring, Field (2011) points out that "second language speakers are even more prone [than first language speakers] to limit their self-monitoring to specific target areas" (p. 85). Self-monitoring is not directly taken account of in the current test. However, aspects of self-monitoring such as self-correction and repair are argued to be captured in the Fluency rubrics. The rubrics also indicate the extent of self-monitoring at different levels.

To address time pressure and reciprocity conditions, each task was analyzed in terms of the time allocated for planning and the total amount of time allowed for completion of the task. In addition, interaction patterns required by each task are presented.

With regard to the second question, the tasks were analyzed in terms of the context validity features outlined in Weir (2005). The task conditions and features are described and evaluated in terms of task setting (response format, task purpose, weighting, known criteria, order of tasks, and time constraints) as well as task demands (channel of communication, discourse mode, length, nature of information, topic familiarity, and lexical, structural, and functional resources). A similar study was done, for example, by Galaczi and ffrench (2011) on Cambridge ESOL examinations. They investigated the Cambridge Main Suite Examinations based on the contextual features noted above to analyze how these feature reflect real-life task settings that are aimed by the tests and the relationship of these features with task difficulty. The current study followed the stages of that study.

With regard to the third question, the parameters related to scoring validity that are proposed by Taylor and Galaczi (2011) are discussed. These parameters include: assessment criteria/ rating scale, rating process, rating conditions, rater characteristics, rater training, post-exam adjustment, and grading and awarding. The

rating scale and assessment criteria are presented in Appendix C. The raters, rating procedures, and rating conditions are described in Chapter 4 to examine the reliability of the rating process. With respect to the reliability of the scores assigned by the raters, the Pearson product-moment correlations among the rater scores were calculated.

The fourth research question was addressed through generalizability (G study) and decision (D study) studies. These analyses were carried out on EduG-6e software program. G study uses the analysis of variance (ANOVA) to partition the variation observed in test scores to reveal the sources of variation (e.g., persons, raters, tasks) and their interactions. It calculates the contribution of each source of variance to the overall score variance. A D study uses the data obtained from the G study to provide estimations about the relative effects of specified numbers of conditions (for example, using one rater and three tasks or two raters and two tasks, etc.) on score dependability.

The test takers were scored on Interaction in Task 3 and Task 6, and they were scored on Coherence in the other tasks. So, in both Form 1 and Form 2, in the third tasks, Coherence was replaced with Interaction; while the other three dimensions were used in all tasks. Therefore, for the G and D studies, Coherence and Interaction scores were pooled into the Coherence/Interaction (C/I) scale.

The decision to use the Interaction rubrics in the tasks that require test taker-test taker interaction, and the Coherence rubrics in the other tasks is supported by Luoma (2004). In her discussion of the adaptation of the CEFR speaking scales for specific speaking assessment contexts, Luoma (2004) points out that “The interaction scale provides some concrete suggestions for wordings when rating interactive skills, while for tasks that require long turns by a single speaker the Coherence scale may

provide some useful concepts" (p. 71). The Coherence scale in the current study relates to the logical organization, development and connection of ideas within the discourse produced by the individual test taker; the Interaction scale, on the other hand, is about a process of joint construction, in which the test taker has to relate his/her utterances to a shared discourse framework. The organization, development, and connection of ideas in the Interaction scale are assessed in relation to the co-constructed discourse; and in this respect, the Interaction rubrics incorporate the Coherence rubrics.

The G and D studies consisted of two phases. The first phase featured a fully crossed p x t x r design (persons crossed with tasks crossed with raters) for the scoring dimensions for each form, Form 1 and Form 2, separately. This means that in the two separate groups, each test taker took each task and was rated by each rater on each task. In this phase, the analyses were based on averaged variance components of all three tasks. The second phase featured a fully crossed p x r design (persons crossed with raters) for each scoring dimension on each of the six tasks. In other words, the relative contributions of persons and raters to overall score variance were investigated, and dependability coefficients for analytic scores at individual task level were obtained. These were used to see which tasks are likely to introduce more error variance in rating a particular scoring dimension. In both phases, D studies were conducted. In these studies, the numbers of the facets (that is, tasks and raters) were varied to examine their impact on the phi coefficients (dependability) of the analytic scores. In other words, these studies showed how score dependability would change when different numbers of tasks and raters were used. Here, the D studies pointed to confidence in scores even in one-task and one-rater condition. To investigate possible problems with this finding, average exact (where the raters gave

the same score), adjacent (where the rater scores differed by 1 point) and nonadjacent (where the rater scores differed by 2 or more points) agreement rates between the four raters were calculated, and individual test taker scores were analyzed through cross-classifications of scores assigned by different raters.

To investigate the fifth research question, the Pearson product-moment correlations among the scores from the dimensions of the analytic scale were calculated by task. The correlations among these analytic scores on individual tasks were found to be very high. However, Xi and Mollaun (2006) argue that even when two sets of scores are perfectly correlated, varied profiles are still likely to exist. To investigate whether the test taker scores varied across the scoring dimensions, the proportion of any two analytic scores differing by 1 or more and by 2 or more was calculated. This was done by comparing each score in the analytical scale given to a test taker in a task with each other. For example, the Fluency score was compared to the scores on Grammatical Range and Accuracy, Lexical Resource, and Coherence/Interaction scores and the resultant differences were grouped as ‘1 or more points difference’ and ‘2 or more points difference’. The differences were calculated for each task and rater. The proportion of the two groups of scores was analyzed for this purpose.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

Weir (2005) discusses two aspects of theory-based validity: a priori and a posteriori validation. This chapter deals with aspects of a priori validation of the test tasks used in the study in terms of their cognitive demands (based on the model proposed by Field, 2011) and contextual features (through a discussion of the aspects of context validity proposed by Weir, 2005, in relation to the test tasks in the current study), and a posteriori validation through a discussion of scoring validity. The first and second research questions, which are about cognitive and context validity, will be discussed below based on the test taker and task characteristics as well as on the rating scale. The third, fourth, and fifth research questions, which relate to scoring validity, will be discussed based predominantly on the statistical analyses of the scores assigned to the test takers. Before these, however, the construct of speaking as is defined in the current study and test taker characteristics will be introduced since they are at the center of test development, and since cognitive and context validation arguments are closely related to these two aspects of test design and development.

In this chapter, detailed analyses of the test tasks in terms of the cognitive demands they place upon the test taker, their contextual parameters, and the scores assigned are presented. In these analyses, frequent references will be made to the tasks and the rating scale. For the final version of the test tasks used in the second pilot examination, and the initial tasks (which were either abandoned or revised following consultation with experts and the first pilot examination), the reader is referred to Appendix A. The task specifications are provided in Appendix B. For the rating scale development process, and the current and previous forms of the rating

scale, the reader is referred to Appendix C. Appendix F presents exemplar performances for different scores on all of the five dimensions of the rating scale and on all of the six tasks used in the present study.

4.2 The construct of speaking in the current test

This section briefly introduces the construct of 'speaking' for the current test, which will be examined in detail in the sections to follow. The current test is aimed at assessing the speaking proficiency of learners of Turkish as an L2 in an academic context and is intended to be used at Boğaziçi University to determine the L2 Turkish speaking proficiency levels of incoming exchange and Erasmus students. It should be noted from the beginning that this is not a test of Turkish for academic purposes, because the general test taker volume do not aim at enrolling in universities in Turkey to get Turkish medium instruction – although there may be a few who plan to do so. Instead, they come to Turkey as exchange or Erasmus students to study at an English-medium university, Boğaziçi University, and at the same time, they would like to develop their proficiency in Turkish as a second (foreign) language. The test, therefore, is a test of general language ability in Turkish, with a particular focus on the uses of language in the university context. It aims at determining these students' current proficiency in Turkish with preliminary alignment with the CEFR levels such that the students can be informed on their language level and be placed at the 'right' Turkish language classes offered at the university.

The test takers come from different backgrounds in terms of their nationalities, native languages, or educational histories. Therefore, they may be considered as a heterogeneous group (See Table 4 and the information about the participants in Chapter 3). However, their experiences with the Turkish language are

similar; most have taken Turkish as a foreign language courses in their universities before coming to Turkey for exchange/Erasmus purposes and/or to develop their proficiency in Turkish. Some, on the other hand, started learning the language after coming to Turkey. They are homogeneous in that they are all university students, and thus, learn and use Turkish mostly in academic environments. However, because they do not get their university education in Turkish, their use of Turkish for academic purposes is limited. Still, because the test takers use the language to communicate with university students around the campus and in classes, the range and functions of their language can be associated to a great extent with the 'campus life'. They need to communicate with instructors and other students on a daily basis on issues related to, for example, their courses, daily activities, holiday plans, events around the campus, as well as more complex topics such as politics, economy, significant events taking place in their home country and Turkey, the subject matter of a course, or decisions taken by the university and so on. Therefore, the 'speaking' skills that the test takers need are mainly shaped by the university environment.

Butler, Eignor, Jones, McNamara, and Suomi (2000) define speaking as "the use of oral language to interact directly and immediately with others" (p. 2). The concern in this test is with the interactions on practical concerns that occur in an academic context. In academic settings, the primary purpose in interactions is acquiring, transmitting, and demonstrating knowledge (Butler et al., 2000). These are related to learning activities, but the test taker population for the current test need to be able to interact with others on issues that do not directly relate to learning or teaching activities such as when they talk with their Turkish friends after school or others in their daily lives. Davies (2001) argues that "sampling for a proficiency test should not be restricted to the work domain", rather it needs to include test tasks that

are relevant to a wider social domain. The current test requires being familiar with university setting in two of the tasks; Task 3 and Task 6. Since the test taker population have gone through schooling and have similar educational backgrounds, these tasks are expected not to be biased towards any sub-group of test takers in terms of the academic knowledge required. The pilot examinations confirmed this expectation, since all of the test takers were able to respond to the content requirements of these test tasks.

Communicative competence in oral academic and everyday language also requires control of a wide range of phonological and syntactic features, vocabulary, and oral discourse, and the pragmatic knowledge of how to use them appropriately to the context and participants, as is put forward in models of communicative competence. In the CEFR (Council of Europe, 2001), for example, communicative competence includes three basic components: language competence, sociolinguistic competence, and pragmatic competence. While the current test incorporates aspects of language competence and pragmatic competence in task design and the assessment criteria, it does not address sociolinguistic competence. In the CEFR, language competence is comprised of lexical, grammatical, semantic, phonological, orthographic, and orthoepic competences. The current test includes Grammatical Range and Accuracy, and Lexical Resource rubrics in the rating scale. Pragmatic competence is divided into three competences in the CEFR: ‘discourse competence’, ‘functional competence’, and ‘design competence’. The current test incorporates aspects of these competences by including various functions in different tasks, and through the Coherence and Interaction rubrics. Knowledge of dialects/varieties or registers, of idiomatic expressions, or of language elements that mark social

relationships, which are aspects of sociolinguistic competence, are not assessed by the current test.

"Success in spoken interaction is determined by (a) the nature of the tasks that the interaction requires and the roles in that interaction; (b) the conditions under which the participants are required to perform; and (c) the resources the individual brings to the interaction" (Butler et al., 2000, p. 2). Therefore, successful performance in the current test will depend on a test taker's ability in the following aspects among others. These aspects are drawn from the CEFR (Council of Europe, 2001) and are assessed in the current test.

- giving descriptions on a range of topics (events, activities, educational background, subjects, etc.),
- developing an argument, highlighting significant points, and giving supporting details,
- giving a talk on a range of topics (a topic within his/her field, complex subjects, etc.),
- participating in routine and non-routine (in)formal discussions,
- exchanging relevant information and giving his/her opinion in daily conversation/meetings/discussions, etc.,
- making appropriate use of important terminology/key lexical items related to the subject matter,
- explaining/passing on main points and detailed information,
- performing and responding to language functions such as information exchange, suggestions, or advice in various situations,
- communicating with reasonable grammatical accuracy,

- using a range of vocabulary for matters connected to his/her field and general topics,
- using a variety of linking words to mark the relationships between ideas,
- demonstrating facility with spoken Turkish while performing the following linguistic tasks:

- summarizing information,
- describing,
- giving and supporting opinions,
- making comparisons /contrasts,
- explaining or informing.

The construct definition will also be relevant in the subsequent sections on the validation arguments for the current test; setting the context of the test should suffice for an introduction.

4.3 Test taker characteristics

In this section, I will consider the test taker characteristics proposed by O'Sullivan (2000, cited in O'Sullivan and Green, 2011; See Table 3) as they relate to the current test.

4.3.1 Physical and physiological characteristics

A key step in addressing the test taker characteristics in test design is to collect data on relevant features. A Learner Profile Form was distributed to test takers in the initial phases of test development. This data gathering procedure provided useful information in obtaining a picture of the profile of the typical test taker population and hence informed the task development process.

Chapter 3 provides the information on 22 of the test takers who took the second pilot test (The data on two candidates were missing as they did not fill in the

Learner Profile Form). The typical age of the test takers is around 21, and this has consequences for test format and content. The task contents in the current test would obviously not be appropriate for younger learners, for example. Although the number of female test takers (16) was more than that of males (6), since learners from both genders sit in the test, the topics and content should reflect the interests of both groups. Moreover, the test takers come from different first language and cultural backgrounds; therefore, it is also important not to privilege a particular L1 or a specific culture. In terms of short-term ailments, such as minor illnesses or injuries, it is advised that the test users make special arrangements, through an estimate of the extent of the disadvantage, such as postponing the exam to a different date. Since the examination will be taken by a limited number of students at Boğaziçi University, this arrangement should not be a major concern for the test providers in the context the test will be used. Longer-term ailments, unfortunately, have not been addressed in the test development process.

4.3.2 Psychological and experiential characteristics

O'Sullivan and Green (2011) argue that any negative effects of personality factors on task performance need to be addressed for a stronger validation argument. A series of other studies have investigated the effect of affective reactions of test takers to various performance conditions (e.g., gender, age and interaction style of the interlocutor/examiner) on their performance (Brown, 2003; Nakatsuhara, 2008; O'Sullivan, 2002; Porter, 1991). In an early study on the effect of examiner gender on test taker performance, the test takers were observed to score higher when interviewed by male examiners (Porter, 1991). Although this study points to gender of the examiner as a significant factor, O'Sullivan and Green (2011) argue that the

results may have been influenced by cultural factors (since the study focused on male Arab students).

O'Sullivan carried out a number of studies investigating the effects of age of the test taker and interlocutor (O'Sullivan, 1995, cited in O'Sullivan, 2002); and acquaintance of interlocutors (O'Sullivan, 2002) on test taker performance in face-to-face speaking test format. He concluded that test takers' performance is influenced by variations in performance conditions and these conditions could result in significant differences in test scores. As mentioned before, similar findings were reported by Brown (2003) and Nakatsuhara (2008), who found effects of interviewers' interaction style on test taker scores.

Experiential characteristics, on the other hand, include educational experience (both in the first language or the target language) and experience in communicating with others, especially in the target language (O'Sullivan and Green, 2011).

Because such variables as examiner gender or experience in the target language can be argued to be components of real life interactions, and the current test aims at assessing interactional competence, trying to eliminate such variables was not the policy adopted in this test development process. Both male and female examiners took part in the current study, and a male and female examiner can be recruited notwithstanding the gender of the test taker when the test is put to operational use. However, the examiners were required to stick to the task administration instructions. Whether the examiners differed in their interaction style and, and if they did, whether this had an impact on test scores were not investigated in the current study.

In the following two sections, I will continue with cognitive demands of the tasks based on Field (2011) and context parameters based on Weir (2005). Readers

are advised to refer to the relevant sections in Chapter 2 (*Literature Review*) for each sub-section below, since the arguments here are built upon the theories, frameworks, and previous studies covered in the literature review. Although references will be made to these studies from time to time, to avoid repetitions they will not be presented here in detail.

Research Question 1: What are the cognitive demands required by the tasks from the test takers?

4.4 Cognitive validity

4.4.1 Conceptualization

Since the current test is intended to cover a broad range of proficiency levels from A2 to C2 in terms of its content in line with the CEFR descriptors of these levels, there appeared to be a requirement to include tasks which differed in how much cognitive demand they impose on the test takers. If, for example, the need for conceptualization were high in all of the test tasks, such that more effortful processing demands were required, then the test takers with limited knowledge and experience in Turkish would potentially find all the tasks difficult, being unable to allocate enough of their working memory on linguistic forms. This would highly likely result in inaccurate sampling/elicitation. As is argued by Field (2011), this range can be accommodated through increasing/decreasing the need for conceptualization and through the support provided to the test taker.

Need for conceptualization: In Task 1, the interview task, the initial questions ask about personal and everyday information, such as the test takers' hobbies and activities. These are basic questions which focus on eliciting personal factual information and on talking about familiar things (e.g., *Boş zamanlarınızda neler yaparsınız?*, *Hangi alanda eğitim görüyorsunuz?*). The questions gradually move

towards non-personal, abstract and therefore conceptually more difficult topics such as the test takers' majors and a comparison between the education systems of two countries. The questions become less predictable, require longer and more engaged responses (e.g., *Favori kitabınız hangisi? Kitapta neler oluyor?*), and invite the test takers to express their opinions on more complex topics (e.g., *Ülkenizin eğitim sisteminde yaşanan en büyük sorunlar nelerdir? Bu konuda diğer ülkeler ile kendi ülkenizi karşılaştırabilir misiniz?*). Task 2 requires the test takers to first describe what they see on the two pictures provided, with no opportunity to deviate from the prompt. Then, they are asked to describe the feelings and reactions of people shown in the pictures and to hypothesize about what might have happened before and what may happen after the events that are portrayed in the pictures. The task, therefore, requires non-personal information and factual information based on concrete input. The questions are less predictable and less familiar. They are semi-controlled in that while the test takers have to stick to the prompt material, the questions include an element of speculation. The need for conceptualization in this task is considered to be moderate. Task 3 is a conceptually more demanding task compared to Task 2. This task, too, includes pictures, but the test takers are expected to go beyond simple description and explain what each picture aims to communicate as well as which picture would best advertise a university. The context in this task is concrete and plausible being within the test takers' personal experience, but it involves abstract concepts based on this concrete context (For example, based on the picture showing students sitting together or playing volleyball on the campus, the test takers are expected to conclude, for example, that this picture indicates the importance the university attaches to social and athletic development of its students). This task is also more open-ended. The test takers are required to evaluate and speculate, express

and justify their opinions, and work towards a negotiated decision. The task, therefore, places high cognitive demands on the test takers.

Task 4 asks the test takers to describe pictures that show different types of holidays and to talk about their holiday preferences, explaining why they would prefer a certain holiday over the others. The task predominantly involves familiar, concrete, and personal information in this regard. The task focuses on giving factual information; however, one question asked within this task requires the test takers to speculate about why the people shown in the pictures chose a particular type of holiday. The test takers also need to make comparisons and contrasts between the holidays to choose one holiday type among the four. The task is considered to impose lower cognitive demands in terms of the nature of information required, but the rhetorical functions required by the one question that requires test takers to speculate can have higher cognitive demands. Task 5 is considered to be the most conceptually challenging task in the current test. In this task, the test takers are required to select one among three topics given to talk on that topic for three minutes after a one-minute preparation time. The questions are non-personal and abstract in nature, and less familiar in terms of content. The test takers are required to go beyond description to exposition and argumentation. And all three questions are open-ended questions. Lastly, Task 6 features a familiar context and requires personal and concrete information. It also requires functions that are familiar and frequent, such as suggesting, asking for information, or requesting facts, and therefore cognitively less challenging. The need for conceptualization in this task is considered to be low.

The support provided: A second consideration, which is proposed by Field (2011) in relation to conceptualization, is how much support is provided to the test takers in the task prompts and by the examiners. This support not only helps in eliciting from test

takers (with different proficiency levels) ratable language, but also structures the tasks so that the language produced can be comparable. In all of the tasks in the current study, the test takers are provided with oral, written, or visual prompts, which may give them ideas that they may wish to express. The visuals also help in limiting the range of language in the prompts that might be reformulated or copied by the test takers.

The tasks include both set questions and more open-ended questions; visual cues that require the test takers to describe, make comparisons, hypothesize, and evaluate; single-, dual-, and multiple-modality prompts; as well as visual, oral, and written prompts. Thus, they present the variety that would be required by a test that aims at different proficiency levels, which would put variable conceptualization demands on the test takers. Field (2011) writes that "An obvious way of grading difficulty … is to gradually reduce this support as the level of the exam increases," (p. 89) which he argues is not the policy adopted in the Cambridge ESOL tests. The tasks used in the current study vary in terms of the information provided to the test takers.

Field (2011) notes that the support provided to test takers can be calibrated by moving:

- from set questions for interlocutors to looser and more open ones
- from precise demands of prompts cards to visual cues that require description and then on to visual cues which need to be compared and evaluated
- from a single prompts (usually in visual form) to multiple ones
- from single-modality prompts to multi-modality ones which combine oral rubrics with both visual and written stimuli. (p. 89)

In Task 1, only oral prompts in the form of questions are used. These move from set questions to more open questions. Task 2 and Task 3 present both oral and visual prompts. However, they differ in the number of visual prompts they provide: two pictures are used in Task 2, while Task 3 features five pictures. Moreover, while the

visual cues in Task 2 require description, those in Task 3 need to be compared and evaluated, and make higher lexical demands. In Task 4, the test takers are provided with four visual prompts. They are expected to describe and make general comments on the pictures. Before the test takers are asked what kind of a holiday they would like to go on, pictures of different holiday types are provided so that they can build their conceptualization on these prompts. In Task 5, the test takers are presented with three questions orally among which they are to select one, and they are asked to talk for three minutes on it. These open-ended questions are also presented to the test takers in written form. However, the test takers are not given any written or visual support that they can use in formulating their responses. Lastly, Task 6 includes oral, written, and visual prompts.

Overall, while the specificity and complexity of the support provided vary from one task to another, it cannot be claimed that this support is systematically calibrated to the presumed CEFR levels of the tasks.

When it comes to 'integrating utterances into a discourse framework', which is another way of operationalizing conceptualization proposed by Field (2011), this is reflected in the interaction tasks (Task 3 and Task 6) and the Interaction rubrics in the rating scale. These two tasks require awareness of the ongoing topic (both Task 3 and Task 6 require following what has been said by the interlocutor and keeping track of details such as which course is appropriate for their interlocutor), recognition of information shared and not shared with the listener (Task 6 is an information gap task), and accommodation to the point-of-view and form of the words of the interlocutor (both Task 3 and Task 6; for example, when the interlocutor expresses a contradictory opinion regarding the best picture to be used in the poster). The Interaction rubrics make explicit reference to test takers' ability to contribute to the

discourse framework at different levels. For example, such rubrics as 'interweave his/her contribution into the joint discourse', 'relate his/her contributions skillfully to those of other speakers', 'indicate when he/she is following', or 'link his/her utterances into clear, coherent discourse' reflect the ability to contribute to the co-constructed discourse (see Appendix C for the rating scale).

4.4.2 Grammatical encoding

Field (2011) points out that the linguistic content of speaking tests is determined by the functions that test takers are required to perform, instead of by grammatical structures. He argues that cognitive demands of a task can be adjusted by reducing/increasing the number of expected functions in relation to a single task and by including functions that are readily available and those that are more difficult to access. He notes that "Functions that are familiar, frequent and concrete will clearly be mapped more rapidly and reliably than those which are not" (p. 91).

Table 6 indicates the functions that are required from the test takers in each task in the current test. They are based on a close investigation of the functions used by the test takers following the second pilot examination.

Table 6. Functional Requirements of the Test Tasks.

Task No	Functional Demands	Example from Test Taker Performances
Task 1	Reporting events (Factual presentation)	*Böyle ilkokuldan başlıyorsun. Dört sene ilk okula gidiyorsun…* *Altmışlarda birçok race sorunlar vardı.* *Buraya geldiğimde bayağı zorlandım en başta.*
	Giving example	*Mesela tarih bir şeyi yanlış anlattığı zaman.*
	Referring to past-present-future	*Uluslararası ilişkiler ve siyaseti istiyordum, onu yazdım kazandım.* *Ders çalışıyorum ve yemeği pişiriyorum.* *Türk üniversitesinde, doğuda çalışacağım.*
	Comparing and contrasting	*Öğrencilerin düşünmesi için gayet yeterli zaman oluyor. Bu bakımda Türkiyeden daha iyi olduğunu düşünüyorum.*
	Giving cause-reason	*Pek hoşuma gitmedi, çünkü İslama karşıydı hem de yani pek sevmedim.* *İlgi alanımdı, o yüzden seçtim.*

	Expressing modality	*Yok ama isterdim.* *Onlar kendine karar verebiliyorlar.*
	Explanation	*Bence yanlış bir şeydir, yani çünkü sonuçta kitaplar hepsi sanattır. Onları hepimiz kendimize göre yorumlarız.*
	Giving an opinion	*Yani özgür bırakılmalı. Başka birinin özgürlüğünü kısıtlamadığı sürece sansür özgür olmalı.*
	Speculating	*Bana kalsa ben üniversitede kalmayı tercih ederim.* *Tam emin değilim. Ya siyaset bilimine ya da uluslararası ilişkilere yöneleceğim.*
	Summarization	*Goodfellasta bir oyuncu var. İsmi Ray Lionel… Hırsız var falan yapıyor. Çok polis onu takip ediyor…*
Task 2	Description (pictures)	*Böyle arkada seyirciler var.* *Tabi ki ben bir maç görüyorum. Yani basketbol. Oyuncular hem de hakem var, başka bir adam var.*
	Reporting events	*Çok sıkı bir yönde konuşuyorlar. Yani o diyor sen geri gitcen o da diyor hiçbir şey yapmadım…*
	Referring to past-present-future	*Deprem yaşanmış. Ev komple yıkılmış.* *Herkes çok üzüldü gibi bakıyor ve herkes yani öyle toprağa bakıyor.* *Yani birbirinize yardım etmeye çalışacaklar.*
	Expressing modality	*Sanırım deprem olmuş olabilir.*
	Speculating	*Belediye illa ki oraları tekrardan yaptıracaktır. Onlarda onu hakemin şeyini Kabul etmemişlerdir.*
Task 3	Description (pictures)	*Yani kütüphaneyi insanlar oturuyor. İşte bilgisayarları laptopları çıkarmışlar.*
	Comparing and contrasting	*Dördüncü fotoğraf daha ilgi çekici olabilir.*
	Suggesting	*Hıhı olabilir. Bence dördüncü resmi seçelim.*
	Giving an opinion	*Bence gelecek öğrencilere üniversitemizin neler sunduğunu göstermesi gerekiyor.*
	Asking for opinion	*Bence dördüncüsü en iyisi. Sen ne dersin? Yani ne düşünüyorsun, başka bir resim sever misin yoksa bunu…*
	Decision making	*-Bence dördüncü resim. -Katılıyorum.* *Karar veremedik.*
Task 4	Description (pictures)	*Dağ ağaçlar var. Gezen birkaç kişi varmış burda.*
	Comparing and contrasting	*İyi bilmiyorum ama daha profesyonel gibi görünüyorlar.*

	Giving cause-reason	*Çok güzel görünüyor, o yüzden sonbaharda dağa gittiler.*
Task 5	Description	*Benim yaşadığım ülke bu arada Hollanda. Bizde çok fazla göçmenler var… Türkler var, Faslar, Polon var.*
	Reporting events	*Okumuyorlar, çalışmıyorlar, hep sokakta dolaşıyorlar.*
	Giving example	*Mesela şimdi Suriyeden gelen insanlar, onlar için çok zor.*
	Referring to past-present	*Onlar ikinci dünya savaşından gelmiş.* *Ev kiralamak çok zor durumunda kalıyorlar.*
	Giving cause-reason	*Tekniksiz insan işçiler doğdu, bu nedenle hep insanları mutlu değil.*
	Expressing modality	*Şu varken bile iş bulabiliyorlar.* *Birleşik şeylerden yardım edilmeli.*
	Speculating	*Eğer anne baba üniversiteye giderse çocuk mutlaka üniversiteye gitmeye düşünecek.*
	Expressing and justifying opinions	*Tabi ki sorunlar var. Bu göçmenler değişik bir kültür, değişik bir dil, değişik bir din Norveçe getiriyorlar.*
Task 6	Requesting facts	*Sen çarşamba programın nasıl?*
	Suggesting	*Belki Günümüzde Avrupa Sineması ilginç olabilir. Ne düşünüyorsun?*
	Comparing and contrasting	*Şekspirden Avrupa Sineması daha ilginç geliyor benim için*
	Giving opinions	*Şekspir biraz sıkıcı görünüyor o yüzden.*
	Asking for opinions	*O zaman Cumhuriyetçilik var. Onu ister misin?*
	Decision making	*Tamam Şekspir okumaları aldık. O zaman bir seçmeli ders aldık.*

As can be seen, the tasks require from the test takers a wide range of functions to be performed, from familiar and concrete functions such as presentation of simple facts to more abstract functions such as speculating and summarizing. The test takers are required in the test, for example, to report factual information about their lives, describe what is happening in a photo, speculate about what might have happened in a particular situation, speculate about what the photos may be communicating about

a university, and explore their usefulness to a poster campaign. This is a clear indication of the range of functions the tasks require.

When it comes to the restriction of functions in relation to a single task, it seems that each task requires multiple functions; however, a closer look at the tasks shows that while each task requires a number of functions, this is done in a step by step progression. For example, in Task 1, different interview questions require different functions, moving from less demanding to more demanding ones. Similarly, in Task 2, the test takers are not required to describe, report, and hypothesize at the same time; quite the contrary, the questions directed by the examiner require one function at a time. The test takers are first asked to describe the pictures, then to report the events shown in the pictures, and lastly to hypothesize about what might have happened before and what may happen after each event. In Task 4, similarly, the test takers are required to give a description first, then to compare the different holiday types, to lastly explain which one they would prefer. In the interaction tasks (Task 3 and Task 6) and the extended talk task (Task 5), however, such a restriction was not possible, since the response requires test taker-test taker interaction without examiner participation – unless one test taker is dominated by the other.

4.4.3 Phonological encoding

Phonological encoding is operationalized in the current test in the Fluency scale. Field (2011) argues that especially at lower levels of proficiency, speakers of a second language will have to pay higher levels of attention and exert more effort while retrieving phonological forms from memory. However, all speakers, irrespective of their level of proficiency, will need to pause to plan, to conceptualize, or to monitor their speech. Therefore, the reasons for dysfluency need to be considered, since the planning process too makes cognitive demands on the test

takers. The Fluency scale for the current test takes some account of the factors that may be responsible for fluency/dysfluency. At higher levels, the reason for pauses may usually be not for processing language, but to plan content. The scale also considers length of run, and duration and frequency of pauses. For speakers with lower proficiency, it is expected that length of run will be short, and pauses will be longer and more frequent. Below are sample scale descriptors of fluency (The list goes from higher levels to lower levels):

‘pauses only reflect search for precision in content’

‘only a conceptually difficult subject can hinder a natural, smooth flow of language’

‘may demonstrate language-related hesitation at times, or some repetition and/or self-correction’

‘even though pausing for grammatical and lexical planning and repair is very evident’

‘with much pausing to search for expressions, to articulate less familiar words, and to repair communication’

4.4.4 Phonetic encoding and articulation

Phonetic encoding and articulation are operationalized by Field (2011) in relation to pronunciation and intelligibility. He argues that “poor L2 pronunciation does not result solely from an inability to form the target sounds but also from an inadequate representation of those sounds in the mind and/or the inability to communicate the appropriate signals to the articulators” (p. 84). Pronunciation does not feature among the assessment dimensions in the current test. Following consultation with two experts, who have been teaching Turkish to foreigners, and analysis of the exam recordings during the assessment scale development, it was decided that pronunciation need not be in the scale, as it was difficult to scale the test takers

according to their pronunciation. Although the test takers at lower levels seemed to have problems with stress and intonation, at higher levels, they did not differ much in terms of intelligibility. Açık (2008) writes that Turkish appears to be an easy language in terms of pronunciation, and attributes the problems learners of Turkish have with sounds such as /ğ/, /u/ and /o/ to their lexicon instead of pronunciation skills. Er, Biçer, and Bozkırlı (2012) list /ğ/, /ş/, /ç/, /ı/ and /ü/ as the sounds which are the most difficult to articulate for learners of Turkish.

Horner (2014), in his evaluation of the CEFR Pronunciation Grid, points out that if the CEFR pronunciation descriptors were formed in terms of intelligibility, they would be as follows:

> C2/ C1/ B2/ B1: Pronunciation is clearly intelligible.
>
> A2: Pronunciation is generally intelligible, but conversational partners will need to ask for repetition from time to time.
>
> A1: Pronunciation of a very limited repertoire of learnt words and phrases can be understood with some difficulty.

Horner (2014) also writes the following:

> Minimally, establishing a hierarchy of phonological elements for each proficiency level ought to be a reflection of research into areas like: which aspects of pronunciation are deemed to be more or less important or irritating (and by whom); or have been shown to have an impact on intelligibility, or to be more or less difficult to acquire, or are acquired in one order rather than another (and to what extent are these last two affected by the learner's L1)?

Other than Açık (2008), and Er, Biçer, and Bozkırlı (2012), studies that investigate the areas of Turkish pronunciation or the learning of Turkish as a foreign language are not available. It seems there is need for further research in this area. However, although the small sample size in the current study makes it difficult to draw sound conclusions, the conclusions seem to be in line with the studies cited above.

4.4.5 Self-monitoring

Even though no explicit reference is made to self-monitoring in the test specifications of the current test, the assessment scale indirectly captures this aspect of speech production. Field (2011) writes that "a competent speaker monitor their own productions for accuracy and appropriacy; and is capable of introducing self-repairs both promptly and following certain norms" (p. 108). Based on this and the discussion of self-monitoring presented in Chapter 3, it seems reasonable to conclude that self-monitoring and self-repair that attend to linguistic features would be evident especially at lower levels of proficiency. This, in turn, may affect the "impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and effectively" (Lennon, 1990, p. 391). This is the definition of fluency as proposed by Lennon. From here, the connection between self-monitoring and fluency can be established. The impression of whether speech is fluent or not would be affected by the self-monitoring and self-repair of the speaker. The following are taken from the Fluency criteria of the current test (Refer to the rating scale in Appendix C). They are ordered from the highest to the lowest level:

'with rare repetition and self-correction'

'with only occasional repetition and self-correction'

'may demonstrate … some repetition and/or self-correction'

'pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production'

'false starts and reformulations are very evident'

'with much pausing … to repair communication'

Moreover, as Skehan (1998, 2009) points out, committing additional cognitive resources to either fluency or accuracy may have a negative impact on the

other. Even though the current test includes Grammatical Range and Accuracy in the rating scale, the relationship between accuracy and self-monitoring is not explicitly established. As is noted by Field (2011), "self-monitoring and self-repair are aspects of speaker performance which are difficult to capture in the form of test specifications" (p. 96).

4.4.6 Time pressure and reciprocity conditions

Two other important variables that influence the cognitive demands of a task are time pressure and reciprocity (Bygate, 1987). In the current test, planning is allowed only in one of the tasks, Task 5, which requires the test takers to choose one of the three topics provided, prepare for one minute, and then talk for three minutes on that topic. The decision to provide planning time for this task was made since the topics that are drawn from the higher levels of the CEFR Speaking Scale descriptors are considered to be conceptually difficult and relatively less familiar, and since only in this task are the test takers required to speak for an extended time without interruption from the examiner, which places heavier conceptualization demands on the test takers. Task 1, the interview task, and the two interaction tasks, Task 3 and Task 6, are interactive tasks that replicate real life speaking events in that they take place under time pressure and require spontaneous production. Moreover, in the interview task the topic changes frequently, and in the interaction tasks the line of conversation is relatively unpredictable (even though both test takers are aware of the topic, it is highly unlikely for a test taker to predict what exactly the interlocutor will say), and thus they place cognitive demands upon the test takers in this respect. Even though the picture tasks as well require a considerable degree of spontaneity, in these tasks the examiners are advised to be patient and wait for a few seconds to allow the test

takers to form ideas about the pictures, retrieve related language (especially vocabulary), and plan what they are going to say.

In terms of reciprocity, the current test employs various formats of interaction and therefore creates reciprocity conditions in the test event. Table 7 presents the interactional patterns and the timings for each task.

Table 7. Interactional Patterns and Timings for Tasks.

Task	Interactional pattern	Minutes
Task 1	E-TT	7
Task 2	TT (E-TT)*	4
Task 3	TT-TT (E-TT-TT)**	5
Task 4	TT (E-TT)*	3
Task 5	TT	3 (+1 min for planning)
Task 6	TT-TT (E-TT-TT)**	5

Note: E = examiner, TT = test taker.
* In this tasks, the examiner asks pre-determined questions and the test taker talks for a period of time without interruption from the candidate; however, the examiner may facilitate talk if the test taker cannot respond at certain points.
** If the interaction is not balanced, such that one test taker dominates the interaction, the examiners may ask questions to level off any such imbalances.

The possible effect of the 'power' difference in the interviewer-test taker tasks are mitigated by the inclusion of tasks that use test taker-test taker format. In addition, the tasks vary in the interaction conditions, as the role of the interlocutor (examiner) and the test taker changes from one task to another; while in the interview task, the length of turn expected from the test takers is short in many of the questions, in picture description tasks they are expected to speak for around one minute without interruption form the examiner for each question. In another task, Task 5, they are expected to talk for three minutes without any interruptions by the examiner. This also accommodates the cognitive demands on the test takers, such that the tasks cover a wide range of difficulty levels in terms of cognitive processing.

Research Question 2: What are the contextual characteristics of the tasks in terms of performance conditions and linguistic demands?

4.5 Context validity

This section addresses the second research question. The contextual parameters of the test tasks are examined based on the aspects of context validity for speaking proposed by Weir (2005). On certain points, the discussion of context validity intertwines with cognitive validity. When this is the case, the readers will be referred in some of the following sections to the relevant section under *Cognitive validity* to avoid repetition.

4.5.1 Test setting

4.5.1.1 Response format

The current test includes various response formats (see Table 8 for the response formats used). This, in turn, provides an opportunity for producing a wider range of speech functions, more production and interaction opportunities, as well as diverse roles for the test takers. The test, therefore, is able to tap into both production (e.g., individual long turn task) and interaction (e.g., tasks that involve discussion between the test takers) skills of the test takers, which is in line with the underlying construct of speaking in the current test, as it has both a cognitive and a social dimension. The use of various response formats contributes to the construct validity arguments, since, as proposed by Galaczi and ffrench (2011):

> The sampling across a range of response formats will broaden the evidence gathered about the test taker's skills, as it will introduce the opportunity for candidates to display a broader range of their communicative competence and will reduce the possibility of construct irrelevant variance. (p. 116-117)

Table 8. Response Formats Used in The Current Test.

Task	Response Formats
Task 1	Interview (short responses for some questions; extended responses for others)
Task 2	Interview (the interlocutor asks questions and the test taker gives extended responses)
Task 3	Test taker-test taker interaction (the two test takers talk to each other)
Task 4	Interview (the examiner asks three questions and the test taker gives extended responses)
Task 5	Individual long turn (the test taker gives a 3-minute talk)
Task 6	Test taker-test taker interaction (the two test takers talk to each other)

4.5.1.2 Task purpose

In the current test, due account has been taken in the design process of clearly communicating the purpose of the tasks to the test takers. For example in Task 1, before moving on to the interview, the examiner states that s/he will be asking a number of questions and expects the test taker to answer those (‘*Sınavın bu bölümü soru-cevap şeklinde olacak. Ben size bazı sorular soracağım. Bu sorulara cevap vermenizi istiyorum.*’); in Task 3, the examiner sets the context for the interaction and asks the test takers to ‘discuss’ and choose the best picture for a poster campaign (‘*Bir üniversite yeni ders yılı için üniversiteyi temsil edecek bir poster hazırlıyor. Bu posterde kullanmak üzere bazı fotoğraflar seçiyorlar. Sizce bu beş resimden hangisi poster için en iyi seçenek olur? Aranızda* tartışın *ve tek bir resim* seçin.’); or in Task 5, the examiner clearly indicates that the task requires extended talk, provides a situation where the test taker is expected to give a talk in a seminar, and informs the test taker that s/he can prepare for one minute before starting to talk (‘*Bir seminerde 3 dakikalık bir konuşma yapacaksınız. Size vereceğim konulardan birini seçin ve bu konu hakkında 3 dakika konuşun. Hazırlık için 1 dakikanız var*.’).

Another important consideration in relation to task purpose is to include different task purposes that may serve the range of proficiency levels included in the test and that may enhance authenticity. Useful categorizations have been outlined in the literature that may serve test developers’ attempt to vary task purposes. Bygate (1987), for example, differentiates between factually oriented (description, narration, instruction, and comparison) and evaluative talk (explanation, justification, prediction, and decision). The CEFR (Council of Europe, 2001) also includes a number of macro-functions (description, narration, commentary, explanation, and demonstration) and micro-functions, such as suggesting, reporting, or greeting. Task

purposes in the form of macro-functions are important, because different functions require different kinds of cognitive processing and may increase/decrease task difficulty (Galaczi and ffrench, 2011). The current test taps into various macro-functions: the test takers are asked to describe (Task 2, Task 4), to narrate (Task 1), to comment (Task 3), to explain (Task 4, Task 6), to justify (Task 4), to compare (Task 3) or to predict (Task 2).

4.5.1.3 Weighting

In the current test, all tasks and all assessment dimensions contribute equally to the overall score assigned to an individual test taker. The test tasks are assessed individually and an overall score is assigned to the test takers on the three tasks according to Fluency, Grammatical Range and Accuracy, Lexical Resource, Coherence/Interaction. Each assessment dimension is equally weighted in their contribution to the overall score. If the need for differential weighting emerges either based on a theory of communicative competence or the program needs at Boğaziçi University, different tasks or rating dimensions may be weighted differently in future administrations of the test.

4.5.1.4 Known criteria

Known criteria relates to whether the test takers are aware of the assessment criteria of the test. The test takers were not informed about the assessment dimensions before sitting in the test in the current study. However, an e-mail was sent to all of the test takers, informing them about the content, structure, and duration of the test tasks. When the test is operationalized, more detailed information can be made available online to the future test takers.

4.5.1.5 Order of tasks

The tasks in the current study could not be ordered as planned in the pilot examinations. The test takers were students who volunteered to take the test in both of the pilot examinations. Since they had a busy schedule and heavy study load, they did not have enough time (and motivation) to take all of the six tasks. Therefore, one half of them took three tasks, and the other half took the other three tasks. So, the participants were divided into two groups, and three tasks were administered to each group (Form 1 and Form 2). As a result, the tasks were not ordered exactly as they had been projected. When the test is put to use, the tasks are planned to be ordered in the following way: Task 1 (Interview), Task 4 (Holiday), Task 2 (Events), Task 5 (Extended talk), Task 6 (Schedule), and Task 3 (Poster). In the pilot examinations too, however, the tasks were ordered according to their assumed difficulty based on the CEFR scale descriptions as far as possible. However, the schedule task in Form 2 disturbed this order, since the paired-task had to be given the last for practical administrative reasons. The order of the tasks as they were given to the test takers in the two forms (Form 1 and Form 2) and the CEFR levels that the tasks were drawn from are presented in Table 9.

Table 9. The Order of Tasks and Their Relevant CEFR Levels.

Order of tasks	Form 1 (CEFR level)	Form 2 (CEFR level)
1	Task 1- Interview (A2 to C2)	Task 4- Holiday (A2 & B1)
2	Task 2- Events (B1)	Task 5- Extended Talk (B2 & C2)
3	Task 3- Poster (B2 & C1)	Task 6- Schedule (A2 & B1)

4.5.1.6 Time constraints

In the current test, five of the tasks do not involve planning time; in only one of the tasks, Task 5, the test takers are allowed to use one minute to prepare, since this task is assumed to be cognitively more demanding, not only because the test takers are required to speak for an extended period (3 minutes) but also because the topics are

conceptually difficult, and relatively less familiar. The other tasks, on the other hand, involve spontaneous speaking, while it should be noted that the examiners are instructed to allow a few seconds for planning in the description tasks so that the test takers familiarize themselves with the pictures. These tasks can be argued to reflect the majority of non-test speaking tasks, which are subject to time constraints and involve spontaneous speaking with no time allowed for planning. Another consideration in the current test was how much time to allocate for each task. The response times were set following the first pilot examination, in which no time constraints were implemented, and based on feedback from the test takers. Hughes (1989) states that "It is unlikely that much reliable information can be obtained in less than 15 minutes, while 30 minutes can provide all the information necessary for most purposes" (p. 105). Even though practicality may arise as an issue, because the current test covers a broad range of proficiency levels, it was decided that the time should be long enough to provide evidence to explicate the range aimed by the test. Practicality issues were to be addressed following validity considerations, through D studies, which will be reported under *Scoring validity*.

4.5.2 Task demands

4.5.2.1 Channel of communication

The approach taken in the current test development is direct testing. The current test aims at enabling inferences about communicative language ability in a general non-test context, and the content is accommodated to conform to the background of the test taking population to facilitate elicitation as far as possible. The direct testing approach enables interaction so that communication is co-constructed by the test takers, and in addition, it allows for extended production in one of the tasks, Task 5, and parts of two other tasks, Task 2 and Task 4. In three of the tasks, Task 2, Task 4,

and Task 3, the test takers are provided with visual prompts, which may provide ample ideas to express, and which are very useful for an independent speaking test.

In the initial test development process, a decision had to be made about whether to use integrated or independent speaking assessment. In integrated tasks, test takers "integrate multiple language skills in a substantial way to complete a speaking task" (Lee, 2006, p. 133). Independent tasks on the other hand, treat speaking as a separate skill (Jamieson, Eignor, Grabe, & Kunnan, 2008). The problems with integrated tasks include, but are not limited to, the fact that they depend on the ability of the test taker in reading or listening and thus it is difficult to attribute the reason of success or failure on the part of the test taker to his/her speaking ability only (Lee, 2006; Jamieson et al., 2008). Because of this problem of relating the inferences drawn from the test to speaking ability, and because the current test was developed as part of a project that included the development of tests of reading, listening, and writing in Turkish as a foreign language, the independent speaking format was favored. Visual prompts were considered to be a good option, since they can provide ideas that the test takers could express, while at the same time limiting the range of language in the prompts that may be reformulated or copied by the test takers. In independent speaking test format, using visuals offered the opportunity to elicit from the test takers a qualitatively and quantitatively wide range of language. However, using visuals only would limit the inferences made from the test to real-world contexts and would not allow for variation (Galaczi & ffrench, 2011). Therefore, different input channels were integrated into the test (for example, aural in the interview, Task 1, and written in the extended talk task, Task 5).

In addition, the language of the prompts has been revised several times based on feedback from experts (three instructors of Turkish as a foreign language) and the

test takers in both the first and the second pilot examinations, so that the level of input language would be within the lexical and structural range of the language level of the test takers. The examiners were also advised to paraphrase in cases of mis- or non-understanding so that even the test takers with low levels of proficiency could understand the requirements of the task and attempt it. At various stages in the test development, feedback was collected and revisions were made to minimize misunderstandings, misinterpretations, or ambiguity. For example, the test takers were observed to misunderstand the question '*Hangi tür kitapları okuyorsunuz*?' and interpreted it as '*Hangi Türk kitaplarını okuyorsunuz?*' So the question was changed to '*Hangi kitapları okursunuz*?' and this seemed to eradicate the ambiguity as the test takers did not have problems in understanding this question in the second pilot examination. Although the same misunderstanding was not observed for the question '*İnternette ne tür konulara ilgi duyuyorsunuz?*', the phrase 'ne tür' was changed to 'hangi'. Moreover, the prompts in Task 3, Task 5, and Task 6 featured relatively longer input lengths. Therefore, the test takers were asked if they had any comprehension problems; if they did, the examiners were advised to repeat or paraphrase the prompt. The task prompts were also provided in the written form to the test takers in Task 5. In this task, the test takers could see the written forms of the questions during the one-minute planning time and while speaking.

4.5.2.2 Discourse mode

Discourse mode is reflected in the current test in relation to the macro-functions identified in the CEFR (Council of Europe, 2001). For a discussion of the discourse modes in the context of the current test, see the section on *Task Purpose* above.

4.5.2.3 Input length

The CEFR (Council of Europe, 2001) indicates that "in general a short text is less demanding than a long text on a similar topic as a longer text requires more processing and there is an additional memory load" (p. 166). Similarly, Galaczi and ffrench (2011) argue that input length has an impact on cognitive complexity. The approach in the current test was to use, as far as possible, short prompts. However, in the interaction tasks, Task 3 and Task 6, and Task 5, the input length is relatively longer, and therefore the language the test takers have to process may impose heavier cognitive demands. In these tasks, due attention was paid not to use complex sentences, and the sentences were kept simple and short. The prompts were revised several times based on expert and test taker feedback. Moreover, the examiners were advised in these tasks to repeat and paraphrase the instructions in case of non-understanding.

4.5.2.4 Nature of information

Galaczi and ffrench (2011) propose investigating nature of information in terms of two criteria: personal/non-personal and abstract/concrete. The initial questions in the interview task, Task 1, are personal, concrete, and familiar ('*Boş zamanlarınızda neler yaparsınız?*', '*Hangi alanda eğitim görüyorsunuz*?' '*Gelecekte nasıl bir işte çalışmak istiyorsunuz?*'), and as the interview progresses, more abstract and less familiar/personal questions are introduced ('*Film ve kitaplara sansür uygulanması hakkında ne düşünüyorsunuz?*', '*Sizce hangi durumlarda sansür uygulanmalıdır?*', '*Yaşlı nüfusun artmasının sizce ne gibi sonuçları olabilir?*'). In one of the picture description tasks, Task 4, the test takers are required to talk about concrete information first (describe holiday types), and then they are asked to hypothesize about the people seen in the photos; so, there is a move form concrete to abstract

information. The other description task, Task 2, is less personal and more abstract in comparison to Task 4. In this task, the test takers are required to speculate about what might have happened before the pictures were taken and what may happen afterwards. Task 6 is an information gap activity, so the test takers do not have knowledge about each other's schedules and have to negotiate to choose two elective courses among a number of courses that are suitable to the schedules of both test takers. This provides an authentic purpose for communication. The setting is assumed to be familiar to the test takers, as they have a background as college students and must have had course schedules in their schools, or college. Task 3 also makes use of visuals. The visuals in this task, however, are not as concrete and familiar as those in Task 4, since the test takers are required to go beyond concrete content in order to speculate about what each picture communicates about the university. This contextual parameter corresponds in substantial ways to Field's (2011) 'conceptualization'. For a more detailed discussion, refer to the section titled *Conceptualization*.

4.5.2.5 Topic familiarity/Content knowledge required

In the current test, one of the tasks, Task 2, appears to be problematic in terms of topic familiarity insofar as the visuals provided to the test takers are concerned. The first picture, which shows a moment from a basketball match where a player and the coach object to a decision made by the referee, may potentially be biased towards (a) candidates from a particular country/culture, and (b) males. It was realized following the second pilot examination that the test takers from the USA knew the teams and players shown in the picture and commented on it accordingly. Since there were only two test takers from the USA, their performance did not feature enough data for qualitative or quantitative analysis. Basketball may also be considered to be more

familiar to male test takers. Qualitative and quantitative analyses into the test taker performances and task scores, however, did not appear to advantage the male test takers; females too could retrieve appropriate vocabulary and speak about the picture as much. As noted earlier, Lumley and O'Sullivan (2005) found only a small effect size for gender in a task on basketball.

Below are parts of the answers given to this question by a few of the female test takers:

'Hatalı bi hareket yaptığını düşündüğü için hakem oyuncuya e ceza vermiş olabilir.'

'Eee birinci resimde adamlar kavga ediyor. Muhtemelen şey sporcu basketbol falan ilgili. Bi tane adam hakem bi tane de oyuncu…'

'Birinci resimde sanırım basketbol sahasında basketbol oynuyorlar hakem yetmiş dokuz numara. Hakem sanırım antrenöre kızıyor, arka tarafta da oyuncu öfkeleniyor.'

Table 10 presents the mean scores assigned to the female test takers on this task (Task 2) by the four raters compared to their mean scores on the two other tasks.

Table 10. Mean Scores of Female Test Takers on Form 1 Tasks.

Test Taker No	Task 2 (Events)	Task 1 (Interview)	Task 3 (Poster)
1	11.88	11.88	11.93
2	2.12	2.31	2.19
3	8.62	8.50	9.00
4	9.38	9.38	9.50
5	10.93	11.38	11.38
6	6.18	6.63	6.25
7	2.50	2.88	2.75
Overall Mean	7.38	7.56	7.57

There were seven female test takers taking Form 1. The overall mean of the female test takers on Task 2 was observed to be lower than their overall mean in the other tasks. Then, the overall mean for the male test takers on Task 2 was calculated. It was 7.28, which was lower than the mean of the females on this task. The overall mean of the five males in Form 1 was found to be 7.43 on Task 1 and 7.58 on Task 3.

These analyses showed that both males and females got lower scores on Task 2 than the two other tasks. This indicated that the females were not penalized in this task.

On the other hand, a qualitative analysis showed that some of the female test takers did not feel 'comfortable' in this task, although this was not reflected in their Task Evaluation Forms. For example, the following were taken from the transcriptions of answers of three female test takers to the picture that described a scene from a basketball match:

'yani basketboldan çok anlamıyorum ama…'

'Ee belki basketboldan pek anlamıyorum ama…'

'hiçbir fikrim yok yani ben ben ee spor falan ilgilenmediğim için…'

These excerpts pointed to problems with this task. The picture, therefore, should be changed with another one, and the task needs to be piloted again before the test is put to operational use.

The second picture in Task 2, on the other hand, seems to be problematic for a different reason. It violates one of the principles listed by Galaczi and ffrench (2011) in relation to Cambridge ESOL examinations, namely that "topic should not be potentially distressing or offensive". The picture which shows a scene after an earthquake may upset a particular sub-group of future test takers; for example, those who have themselves experienced an earthquake. Even though the test takers in the second pilot examination did not react in such a way to the picture (this picture was included in the task to replace another picture that was also found to be potentially biased in the first pilot examination), this does not preclude the possibility that it will be distressing to some test takers in future examinations.

Apart from these, the topics in the test elicited a wide range of language in terms of quantity as well as quality (functions, grammatical and lexical range, etc.), which is discussed below.

4.5.2.6 Lexical resources, structural resources, and functional resources

In Appendix G, the number of the types and tokens of the words used by the test takers in four of the tasks, Task 1, 2, 4, and 5, are presented (The interaction tasks were not included in this analysis since the distribution of talking time was not always similar for the two interlocutors in these tasks). There was not discovered a linear relationship between the number and types of words used by a test taker and his/her score on the Lexical Resource scale. A similar finding was reported by Davis (2009), who found that the test takers who produced more amount of language did not necessarily obtain higher scores. This may be attributed to the 'precision', 'accuracy' and 'appropriateness' criteria noted in the scale. Moreover, every word used by the test takers was included in the word count; therefore, repetitions ('*ama e:::m e:: ama*', Brian; '*ne zaman a::: ne zaman ben… diğer bir şey e::: diğer şey e::: bir şey e:: bir şe:y*', Jonathan) and fillers such as 'şey' and 'yani' contributed to the count, even though these test takers were in fact penalized for hesitations and obvious searching for vocabulary. Natalie, for example, used the word 'yani' 73 times in the interview task; the total number of the tokens she used was 616. She was not assigned a higher score by the raters for using this particular word, but the word was included in the word count. The discrepancy, therefore, can be accounted for by "goodness/appropriacy of usage" (Schmidt, 2009, cited in Galaczi & ffrench, 2011), which was rewarded by the raters over simply using a certain word or structure. That being said, it must be pointed out that even though there was not a one-to-one correspondence between the scores and number/type of tokens, there was evidence

that an increase in level was associated with an increase in the number of word types and tokens. As is noted by Davis (2009), “higher word count per se does not appear to generate higher scores; rather, word count and score appear to be jointly correlated to a third variable, such as general oral language proficiency” (p. 387).

When it comes to structural resources, the current test assesses the test takers’ structural resources through the Grammatical Range and Accuracy criteria in the rating scale. The scales are very close to those identified in the CEFR (Council of Europe, 2001), as the rating scale development process was informed by the CEFR (For more on the rating scale, see Appendix C). Lastly, for a discussion of functional categories in relation to the current test, see the section titled *Grammatical encoding* above.

4.6 Scoring validity

In this section the third, fourth, and fifth research questions will be addressed. With regard to the third question, I will briefly discuss the parameters related to scoring validity that are proposed by Taylor and Galaczi (2011) and provide inter-rater reliability coefficients. The parameters Taylor and Galaczi (2011) propose are: assessment criteria/ rating scale, rating process, rating conditions, rater characteristics, rater training, post-exam adjustment, and grading and awarding. The rating scale and assessment criteria are presented in Appendix C, and due discussion on the rationale has been presented above. Here, I will focus on the ratings and the raters. Then, I will continue with statistical analyses that have been carried out to investigate the dependability of the test scores assigned to the test takers in the current study and the seperability (distinctness) of the scoring dimensions based on the test scores, which are the focus of the fourth and fifth research questions,

respectively. Post-exam adjustment and awarding will not be discussed as these issues are beyond the scope of the present study.

Research Question 3: How reliable are the rating process and the scores assigned to the test takers?

4.6.1 The rating process

4.6.1.1 Rating and raters

The rating scale and the steps involved in the adaptation of the scale can be found in Appendix C. The rating process is about the decision-making processes that the raters go through (Taylor & Galaczi, 2011). In order to simplify this process for the raters, the CEFR (Council of Europe, 2001) suggests that the scale descriptions should be positive, definite, clear, independent, and brief. Pollitt and Murray (1996), similarly, point to the need to make this process simple and transparent. Since the rating scale was adapted from the CEFR scale descriptors, it was assumed that the scale descriptors would provide clear and definite descriptions, helping ease the scoring process for the raters. Rater perceptions of the descriptors were not explicitly investigated in the current study.

In terms of rating conditions, Taylor and Galaczi (2011) emphasize the temporal, spatial, and psychological dimensions of different rating conditions. In the current study, the raters did not have to make real-time judgments; instead, they had the opportunity to listen to the performances several times on the recordings before arriving at a decision about the test taker performances. Because the raters were not involved in the test administration, and because the performances were not video-taped, they did not have the chance to observe the test takers' non-verbal communication behavior. However, since non-verbal communication skills were not

defined as a part of the construct of speaking ability in the current test, it did not feature as a matter of concern.

Myford and Wolfe (2004) list a number of possible rater effects that may threaten the scoring validity of a test: (a) rater leniency or severity, (b) the halo effect (for example, assigning all analytic scores based on an overall impression), (c) inconsistency in applying the analytic dimensions, and (d) rater bias against a sub-population of test takers. These aspects of the rater characteristics were addressed through rater training and rater monitoring.

The raters were native speakers of Turkish. They consisted of four senior undergraduate students of Foreign Language Education Department, Boğaziçi University. They did not have experience in scoring L2 Turkish performances. A rater training session on the rating scale and scoring procedures took place before the scoring. The session lasted for a total of around six hours. Student performances from the first and second pilot tests were used to exemplify performances at different levels for each dimension in the rating scale. Following the ratings, the scores were compared by the researcher and when discrepancies were observed, the raters were asked to rate for a second time these particular performances on which there were disagreements. The raters were free to change the scores they assigned in the first rating or give the same score again. In a number of cases, they decided to keep the first scores, while in others they assigned a different score. The goal in rater training and monitoring was to minimize differences in the ratings. However, as was reported by Hubbard, Gilbert, and Pidcock (2006), even after rater training, raters may focus on different aspects of performance in different parts (tasks) of the test. In the case of the current study, this does not seem very likely to have happened; since, unlike in the CAE Speaking test, which was the focus of Hubbard et al. (2006) study, in the

current test, raters assigned scores to each individual task instead of giving an overall score for the whole test (Note that, the rating scale of CAE Speaking test also uses an analytic scale). In other words, the raters had to give separate analytic scores on the four scoring dimensions on each of the three tasks in both forms, and thus they were required to focus on all dimensions on all tasks. An investigation into rater behavior and the reasons why they assigned a particular score to a particular performance, through verbal protocol for example, could have provided valuable information; but, this was not possible within the scope of the current study.

4.6.1.2 Inter-rater correlations

The 12 test takers in each form were rated on each of the five dimensions of Fluency (F), Grammatical Range and Accuracy (A), Lexical Resource (L), Coherence (C) and Interaction (I) by four raters. Refer to Appendix H for the scores assigned to the test takers by the four raters. Table 11 and Table 12 provide the mean scores of the test takers (all names are pseudonyms) in Form 1 and Form 2 respectively, as well as the overall means in the two forms and standard deviations. The tables show that in both forms, there were test takers with different levels of proficiency, and they formed a heterogeneous group in this respect.

Table 11. Mean Scores of the Test Takers in Form 1

Test Taker	Mean Score (assigned by the 4 raters)
Eleanor	2.21
Nola	2.71
Jonathan	4.02
Ryan	5.56
Brian	6.29
Olivia	6.35
Natalie	8.71
Anna	9.42
Douglas	9.73
Christina	11.23
Andrew	11.56
Barbara	11.90
Overall Mean	7.47
SD	3.43

Table 12. Mean Scores of the Test Takers in Form 2.

Test Taker	Mean Score (assigned by the 4 raters)
Mindy	3.29
Alphonso	3.56
Fabio	5.45
Lee	5.83
Kurt	5.83
Rebecca	6.17
Gilbert	6.58
Grace	7.54
Kathy	8.77
Amanda	9.42
Clint	9.60
Sona	11.08
Overall Mean	6.92
SD	2.41

Table 13 and Table 14 show the correlations between rater pairs by the five dimensions by tasks for both forms, Form 1 (Task 1, 2 and 3) and Form 2 (Task 4, 5 and 6). All of the correlation indices were substantially high and significant. This indicates that the test takers were rank ordered similarly in all of the five dimensions by different rater pairs. However, these results need to be interpreted tentatively both because of the problems associated with correlation indices in CTT (for example, rater scores may be highly correlated but that does not preclude the possibility that the raters function differently in terms of leniency or severity), and because the sample size was small and the group of test takers were heterogeneous in terms of speaking ability as can be seen on Table 11 and Table 12. Moreover, the raters were asked to rate the test takers a second time when discrepancies were observed among the raters in the first ratings. Thus, it is not surprising that very high inter-rater correlations were obtained considering the small sample size, the heterogeneity of the group of test takers, and the training and monitoring of the raters. However, as will be pointed out in the following sections, despite the high correlations, there still were important discrepancies between the scores assigned by the raters, even though these were few in number. The details and implications of this are discussed in the following sections.

Table 13. Correlations among Raters by Scoring Dimension by Task for Form 1.

	R1 & R2	R1 & R3	R1 & R4	R2 & R3	R2 & R4	R3 & R4
T1 F	.977**	.982**	.994**	.979**	.980**	.984**
T1 A	.973**	.978**	.994**	.942**	.970**	.985**
T1 L	.961**	.984**	.979**	.937**	.985**	.974**
T1 C	.987**	.992**	1.00**	.992**	.987**	.992**
T2 F	.989**	.985**	.994**	.987**	.984**	.988**
T2 A	.956**	.978**	.973**	.931**	.955**	.981**
T2 L	.970**	.983**	.985**	.969**	.983**	.970**
T2 C	.978**	.984**	.994**	.970**	.983**	.979**
T3 F	.979**	.977**	.981**	.979**	.980**	.973**
T3 A	.989**	.964**	.990**	.980**	.986**	.965**
T3 L	.989**	.971**	1.00**	.981**	.989**	.971**
T3 I	.995**	.988**	.995**	.989**	.993**	.989**

Note. T1 = Task 1, T2 = Task 2, T3 = Task 3; F = fluency, A= grammatical range and accuracy, L = lexical resource, C = coherence, I = interaction; R1 = rater 1, R2 = rater 2, R3 = rater 3, R4 = rater 4.
* $p < .05$, ** $p < .001$.

Table 14. Correlations among Raters by Scoring Dimension by Task for Form 2.

	R1 & R2	R1 & R3	R1 & R4	R2 & R3	R2 & R4	R3 & R4
T4 F	.980**	.927**	.975**	.930**	.959**	.942**
T4 A	.993**	.959**	1.00**	.966**	.993**	.959**
T4 L	.943**	.926**	.941**	.958**	.952**	.958**
T4 C	.987**	.964**	.988**	.970**	.982**	.970**
T5 F	.994**	.975**	1.00**	.980**	.994**	.975**
T5 A	.987**	.964**	1.00**	.975**	.987**	.964**
T5 L	.976**	.974**	.974**	.992**	.980**	.985**
T5 C	.971**	.984**	.994**	.979**	.977**	.985**
T6 F	.976**	.970**	.977**	.977**	.983**	.981**
T6 A	.988**	.964**	.989**	.976**	.976**	.954**
T6 L	.972**	.988**	.986**	.972**	.977**	.986**
T6 I	.987**	.990**	1.00**	.983**	.987**	.990**

Note. T4 = Task 4, T5 = Task 5, T6 = Task 6; F = fluency, A= grammatical range and accuracy, L = lexical resource, C = coherence, I = interaction; R1 = rater 1, R2 = rater 2, R3 = rater 3, R4 = rater 4.
* $p < .05$, ** $p < .001$.

Research Question 4.1. To what extent are the analytic scores dependable?

Research Question 4.2. How many tasks and raters would produce relatively more dependable scores?

4.6.2 G study, D study, rater agreement and cross-classifications for score dependability analysis

These statistical analyses provide an empirical investigation into the dependability of the analytic scores assigned to the test takers and into how score dependability changes when different numbers of tasks and raters are used (Xi & Mollaun, 2006). An important aspect of this phase is that it provides an in-depth analysis of the scores

going beyond the results obtained from generalizability and dependability analyses to look into the individual test taker through cross-classifications of the scores assigned to an individual test taker by different raters. These analyses aimed to find the most effective test design in terms of the number of tasks and raters to be used.

Table 15 shows the means and standard deviations of the analytic scores of the 24-person sample by task.

Table 15. Descriptive Statistics for the Analytic Scores.

	Dimension	M	SD
Task 1	F	7.47	3.60
	A	7.37	3.19
	L	7.27	3.30
	C	7.91	3.16
	Total	7.51	3.40
Task 2	F	7.41	3.53
	A	7.08	3.37
	L	7.08	3.28
	C	7.77	3.34
	Total	7.33	3.46
Task 3	F	7.35	3.57
	A	7.54	3.15
	L	7.29	3.26
	I	8.10	3.35
	Total	7.57	3.42
Task 4	F	6.52	2.22
	A	6.77	2.31
	L	6.81	2.13
	C	7.54	2.27
	Total	6.91	2.25
Task 5	F	6.58	2.51
	A	6.47	2.18
	L	6.33	2.14
	C	7.16	2.31
	Total	6.64	2.34
Task 6	F	7.00	2.55
	A	7.14	2.65
	L	7.06	2.49
	I	7.72	2.60
	Total	7.23	2.65

Note. F = fluency, A= grammatical range and accuracy, L = lexical resource, C = coherence, I = interaction; M = mean, SD = standard deviation.

The means of different dimensions and of different task and dimension combinations were quite close. In tasks 4 and 5, the mean scores were relatively lower than the means of the other tasks; however, the difference was not substantial. Overall, the means of the tasks in Form 2 were less than those in Form 1. This may be an indication of task difficulty; however, since there were only a small number of test

takers taking each form, this difference may be a result of the difference between the proficiency levels of the test takers in each form.

4.6.2.1 Univariate G analyses on the dependability of analytic scores

In these analyses, Coherence and Interaction scores were pooled into the Coherence/Interaction (C/I) score since the test takers were scored on Coherence in the first two tasks in both forms, and Coherence was replaced with Interaction in the third tasks. The first phase of the present statistical study featured a fully crossed p x t x r design for the four scoring dimensions for both forms, Form 1 and Form 2. The second phase featured a fully crossed p x r design for the four scoring dimensions for each of the six tasks. This second phase will be discussed below under *Dependability of Analytic Scores by Task*.

Table 16 and Table 17 show the variance components and the percentages of the total variance accounted for by each source of variance. The analyses were carried out for each scoring dimension separately to investigate score dependability for each of the analytical dimensions. The F, A, L, and C/I columns under 'Variance component' present variance component values associated with sources of variance (persons, tasks, and raters) and their interactions (person-by-task interaction, person-by-rater interaction, task-by-rater interaction, and person-by-rater-by-task interaction plus undifferentiated error) for Fluency, Grammatical Range and Accuracy, Lexical Resource, and Coherence/Interaction dimensions, respectively. The 'Percent of total variation' columns show these variance component values converted to a percentile scale for ease of interpretation. These columns, therefore, present the proportion of variance of individual scores that is attributable to each variance source. Xi and Mollaun (2006) explain the interpretation of each source of score variance as follows: Persons (or test takers) constitute the object of measurement, not error, and

indicate systematic individual differences in test taker ability. The other sources of variation are considered sources of error. Task main effect (T) indicates differences in difficulty levels of the tasks; so, if the proportion of variance attributed to tasks is large, that means the test takers' scores differed across the tasks. The rater main effect (R) indicates raters' leniency or harshness in their ratings, that is, the extent to which the mean scores assigned by different raters to the same performance are the same. Large rater main effect can be interpreted as differences in rater judgments. Interaction effects, on the other hand, indicate (in)consistency in the rank ordering of test takers. A large person by rater interaction, for example, indicates that the test takers are rank ordered differently by different raters.

Table 16. Variance Components for the p x t x r Design. (Form 1)

Sources of variation	Variance component				Percent of total variation			
	F	A	L	C/I	F	A	L	C/I
P	13.383	10.970	11.351	11.607	96.4%	95.1%	95.9%	97.5%
T	.005	.047	.013	.011	0.0%	0.4%	0.1%	0.1%
R	.149	.101	.101	.006	1.1%	0.9%	0.9%	0.1%
PT	.056	.007	.005	.046	0.4%	0.1%	0.0%	0.4%
PR	.128	.202	.166	.086	0.9%	1.8%	1.4%	0.7%
TR	.007	.004	.004	.014	0.1%	0.0%	0.0%	0.1%
PTR	.163	.196	.198	.129	1.2%	1.7%	1.7%	1.1%

Note. P = person, T = task, R = rater, PT = person by task, PR = person by rater, TR = task by rater, PTR = person by task by rater; F = fluency, A = grammatical range and accuracy, L = lexical resource, C/I = coherence/interaction.

Table 17. Variance Components for the p x t x r Design. (Form 2)

Sources of variation	Variance component				Percent of total variation			
	F	A	L	C/I	F	A	L	C/I
P	6.204	5.936	5.284	6.011	94.7%	93.7%	92.4%	95.2%
T	.055	.097	.118	.063	0.9%	1.5%	2.1%	1.0%
R	.011	.001	.000	.037	0.0%	0.0%	0.0%	0.6%
PT	.083	.148	.063	.042	1.3%	2.3%	1.1%	0.7%
PR	.044	.005	.013	.000	0.7%	0.1%	0.0%	0.0%
TR	.018	.002	.033	.027	0.3%	0.0%	0.6%	0.4%
PTR	.147	.146	.220	.129	2.3%	2.3%	3.9%	2.1%

Note. P = person, T = task, R = rater, PT = person by task, PR = person by rater, TR = task by rater, PTR = person by task by rater; F = fluency, A = grammatical range and accuracy, L = lexical resource, C/I = coherence/interaction.

Table 16 and Table 17 indicate that the test takers were almost equally variable in their fluency, accuracy, lexical resource, and coherence/interaction in both forms, as is indicated by the similar variance components associated with persons (true variance in CTT) on these four dimensions. That is, the variance associated with the persons (test takers) were similar for all of the four dimensions. Overall, a substantial proportion of the total variance in scores on the four dimensions could be explained by real differences in the test takers' fluency, accuracy, lexical resource, or coherence/interaction. Among the four dimensions, 97.5% in Form 1 and 95.2% in Form 2 of the variance in the test takers' coherence/interaction scores was explained by variance associated with persons, suggesting that the test takers' scores in this dimension were the most reliable; yet, it should be noted that the percent of variation explained by real differences in person ability was over 90% in all task-by-dimension combinations.

These results suggest that the test takers were rank ordered similarly on each of the four dimensions by the four raters; so, the raters did not differ much in judging where a test taker stood compared to the other test takers. In other words, in only a few cases, the test takers were rank ordered differently by the raters. The small rater main effects indicate that the raters did not differ much in their leniency or harshness. The task main effects and the person-by-task interaction were almost zero, indicating that the mean scores of this group of test takers were similar across the tasks (i.e., on average, the tasks varied little in their difficulty levels) and the scores assigned to the test takers were similar on each of the four dimensions across the tasks. The task-by-rater interaction was negligible too, pointing to minimal difference in the raters' rank orderings of task difficulty.

However, these results need to be approached carefully in order not to reach possible erroneous interpretations. The G coefficient will be higher when the sample size is small and when the target population is more heterogeneous (Xi & Mollaun, 2006). The number of test takers in the two forms was not very large (12 test takers took each form), so the sample size was small. The test takers in both forms were from a range of different proficiency levels and hence formed a heterogeneous group in terms of proficiency. Moreover, the scores were assigned following rater monitoring, which attempted to minimize discrepancies among the raters.

Overall, the G analyses indicated that the raters did not differ much in their leniency or harshness. This may be attributed to rater training and monitoring or to small group sizes, or it may be a result of the heterogeneity of the groups. The tasks were expected to be of different difficulty levels; however, the G analyses suggested the opposite, that is, the tasks were found to be of similar difficulty for this group of test takers.

4.6.2.2 D studies: Changes in phi coefficients

In the G studies, the contributions of different sources of variation to score variance were estimated. A D study follows a G study. In the D study, "the values determined for the estimated variance components [in the G study] are used to further calculate estimates of the effects of various measurement designs on the *dependability* (analogous to *reliability*) of the scores" (Brown, 2005, p. 13). In the current study, D studies were conducted where the number of tasks and raters were varied to examine their impact on the phi coefficients[4] of the analytic scores (Xi & Mollaun, 2006). The results of the D studies were expected to offer useful information about optimizing

[4] The g- and phi-coefficients reported in generalizability analyses can both be interpreted as reliability coefficients. They describe how well the mean rating for an individual examinee predicts her universe score (Sudweeks, Reeve and Bradshaw, 2005, p. 244).

the assessment design by showing how many tasks and raters would produce more dependable results. Since the dependability and validity of the assessment was the immediate concern in the test development, drawing upon Weir's (2005) argument that "practicality is simply not a necessary condition for validity. Only when sufficient evidence is available to justify interpreting test scores as an acceptable indication of the control of an underlying construct should we concern ourselves with practicality" (p. 49), practicality did not feature as one of the basic considerations in the initial task development process in the present study. However, efficiency and practicality of assessment also need to be considered when a test is to be put to operationalized use. D studies provide valuable information in this regard.

Table 18 and Table 19 provide the phi coefficients for the analytic scores when different combinations of number of tasks and raters are used for a fully crossed p x T x R design for both Form 1 and Form 2. In these tables, the F, A, L, and C/I columns under 'single rating' show what the phi coefficient values would be if test taker performances were rated only once with different numbers of tasks used (from 1 to 6 tasks). For example, for Fluency, with one rating and one task, the obtained phi coefficient would be .963, while it would increase to .976 with one rating and six tasks. This indicates that using six tasks produces more dependable scores. However, .963 is already indicates high dependability. The columns under 'double rating' show how the phi coefficients would change if the performances were rated twice with different numbers of tasks.

The tables show that the phi coefficient is substantial even when one task and one rater are used since, as was observed in the G study results, the tasks did not differ much in their difficulty levels, the test takers obtained similar scores across the tasks, and the raters awarded consistent scores across the test takers and tasks.

Table 18. Changes in Phi Coefficients of the Four Analytic Scores in D Studies. (Form 1)

	Alternative D studies for p x T x R design								
		Single rating				Double rating			
	# of tasks	F	A	L	C/I	F	A	L	C/I
Phi coefficient	1	.963	.951	.958	.975	.979	.972	.978	.985
	2	.971	.962	.967	.983	.984	.979	.983	.990
	3	.974	.965	.970	.986	.986	.981	.984	.992
	4	.975	.967	.972	.987	.987	.982	.985	.993
	5	.976	.968	.973	.988	.987	.983	.986	.993
	6	.976	.969	.973	.989	.988	.984	.986	.994

Note. F = fluency, A = grammatical range and accuracy, L = lexical resource, C/I = coherence/interaction.

Table 19. Changes in Phi Coefficients of the Four Analytic Scores in D Studies. (Form 2)

	Alternative D studies for p x T x R design								
		Single rating				Double rating			
	# of tasks	F	A	L	C/I	F	A	L	C/I
Phi coefficient	1	.946	.937	.923	.952	.962	.948	.944	.967
	2	.969	.966	.960	.972	.978	.973	.971	.981
	3	.976	.977	.973	.979	.984	.981	.980	.986
	4	.980	.982	.979	.983	.987	.986	.985	.989
	5	.983	.985	.983	.985	.989	.988	.988	.990
	6	.984	.987	.986	.986	.990	.990	.990	.991

Note. F = fluency, A = grammatical range and accuracy, L = lexical resource, C/I = coherence/interaction.

One obvious observation from the D studies was that the phi coefficients increase when more raters and more tasks are used. When one rating is obtained for each task, using two or three tasks yields higher phi coefficients than with one task, but as the number of tasks increases from four to six, the improvement in phi coefficients is less dramatic.

The analyses above illustrate how the phi coefficients would change with different combinations of number of tasks and raters. The analyses are based on averaged variance components for all three tasks in each form. Information on the dependability of analytic scores at the task level may also be useful since it shows which tasks may introduce more unreliability in scoring a particular dimension (Xi & Mollaun, 2006). This is the focus of the next section.

4.6.2.3 Dependability of analytic scores by task

At this stage, each task was analyzed separately for variance components for the four scoring dimensions. Each analysis featured a fully crossed p x r design. The variance components associated with different sources of variation for different tasks are shown in Table 20.

Table 20. G Study Variance Components by Tasks.

		Variance component				Percent of total variation			
		F	A	L	C/I	F	A	L	C/I
Task 1	P	13.703	10.713	11.393	10.766	96.4%	96.0%	95.9%	98.4%
	R	.253	.093	.093	.010	1.8%	0.8%	0.8%	0.0%
	PR	.253	.351	.399	.176	1.8%	3.1%	3.4%	1.6%
Task 2	P	13.361	11.690	11.217	11.902	98.1%	94.0%	95.4%	97.3%
	R	.002	.179	.122	.053	0.0%	1.4%	1.0%	0.4%
	PR	.252	.570	.419	.273	1.9%	4.6%	3.6%	2.2%
Task 3	P	13.417	10.531	11.327	12.056	96.3%	97.1%	97.4%	98.2%
	R	.171	.044	.075	.012	1.2%	0.4%	0.7%	0.1%
	PR	.349	.275	.229	.202	2.5%	2.5%	2.0%	1.7%
Task 4	P	5.118	5.685	4.573	5.444	94.9%	97.6%	92.0%	96.8%
	R	.002	.010	.046	.011	0.0%	0.0%	0.9%	0.2%
	PR	.273	.142	.349	.169	5.1%	2.4%	7.0%	3.0%
Task 5	P	6.782	5.102	4.864	5.577	98.4%	97.7%	97.0%	95.3%
	R	.000	.006	.051	.161	0.0%	0.1%	1.0%	2.8%
	PR	.111	.111	.101	.116	1.6%	2.1%	2.0%	2.0%
Task 6	P	6.936	7.465	6.606	7.291	97.3%	97.4%	97.4%	98.6%
	R	.032	.001	.001	.022	0.5%	0.0%	0.0%	0.3%
	PR	.161	.202	.172	.081	2.3%	2.6%	2.5%	1.1%

Note. P = person, R = rater, PR = person by rater; F = fluency, A = grammatical range and accuracy, L = lexical resource, C/I = coherence/interaction.

Overall, the variances explained by sources of error were very small. The task with the largest source of error associated with person-by-rater interaction is Task 4, followed by Task 2. These are the two picture description tasks used in the current test. Task 4 is the holiday task, where test takers are asked to describe different types of holidays shown in the pictures and then choose one type of holiday, explaining why they would prefer that particular holiday type over the others. Task 2 is the other picture description task, which asks the test takers to describe two events shown in the pictures and then hypothesize about what might have happened before each picture was taken and what may happen afterwards. However, it needs to be noted

again that the error variance in all of the tasks was very small, including these two picture description tasks.

Table 21 shows the results of the D studies, where phi coefficients were compared for single versus double raters for a p x R design. Substantially high phi coefficients were obtained in all task-dimension combinations and the phi coefficients were very close in all combinations. One obvious observation is that the phi coefficient increases when two raters are used.

Table 21. Changes in Phi Coefficients by Task.

		Single rating				Double rating			
		F	A	L	C/I	F	A	L	C/I
Phi coefficient	Task 1	.981	.960	.958	.983	.990	.979	.978	.991
	Task 2	.981	.939	.953	.973	.990	.968	.976	.986
	Task 3	.962	.970	.973	.982	.980	.985	.986	.991
	Task 4	.949	.975	.920	.967	.973	.987	.958	.983
	Task 5	.983	.977	.969	.952	.991	.988	.984	.975
	Task 6	.972	.973	.974	.985	.986	.986	.987	.992

Note. F = fluency, A = grammatical range and accuracy, L = lexical resource, C/I = coherence/interaction.

4.6.2.4 Rater agreement

Table 22 shows the rater agreement rates by dimension and by task. The rater agreement rates were higher for the tasks in Form 2 (Task 4, 5 and 6). The combined agreement rates were similar for all of the four dimensions. The overall nonadjacent discrepancy rate for Coherence/Interaction was the lowest among the four dimensions, indicating that the raters had the largest agreement rate for this dimension. The nonadjacent agreement rates for Lexical Resource, Grammatical Range and Accuracy, and Fluency were slightly higher. Some of the largest percentages of nonadjacent discrepancies occurred with Fluency in Task 1 and Task 3, Grammatical Range and Accuracy and Lexical Resource in Task 2. Most of these nonadjacent discrepancies were mainly associated with two rater pairs, Rater 2 and Rater 3, and Rater 3 and Rater 4, which accounted for 30 of 50 discrepancies. This raises questions about including only one rater in the assessment process, even

though the D studies produced high dependability coefficients for one rater. Moreover, it should be noted that these results were obtained following rater monitoring, where the aim was to remove the discrepancies by asking the raters to rate for a second time the performances for which discrepancies were observed.

Table 22. Average Agreement Rates between the Four Raters on Each Scoring Dimension by Task.

Task	Fluency			Gram. Rang. & Acc.			Lexical Resource			Coherence/Interaction		
	Exact (%)	Adj. (%)	Non. (%)	Exact (%)	Adj. (%)	Non. (%)	Exact (%)	Adj. (%)	Non. (%)	Exact (%)	Adj. (%)	Non (%)
1	44.4	40.3	15.3	51.4	37.5	11.1	56.9	31.9	11.1	66.7	33.3	0.0
2	50.0	50.0	0.0	43.1	36.1	20.8	48.6	34.7	16.7	50.0	44.4	5.6
3	33.3	50.0	16.7	63.9	27.8	8.3	65.3	26.4	8.3	52.8	47.2	0.0
4	54.2	43.1	2.8	72.2	27.8	0.0	41.7	48.6	9.7	63.9	36.1	0.0
5	77.8	22.2	0.0	76.4	23.6	0.0	69.4	30.6	0.0	52.8	44.4	2.8
6	61.1	38.9	0.0	59.7	40.3	0.0	65.3	34.7	0.0	79.2	20.8	0.0
Total	53.5	40.7	5.8	61.1	32.2	6.7	57.9	34.5	7.6	60.9	37.7	1.4

Note. Adj. = adjacent, Non. = nonadjacent.

4.6.2.5 Cross-classifications of the ratings

All of the analyses presented so far provided summary statistics for raters, tasks, criteria, or the interactions between them. Despite important discrepancies between two rater pairs, overall high agreement between the four raters was obtained. The generalizability and decision studies also produced high generalizability and dependability coefficients, pointing to confidence in scores even with a one-rater condition. However, a closer investigation into rater behavior points to problems with this interpretation, indicating that even very high indices of inter-rater correlation or high dependability coefficients may not justify using one rater only. Cross-classifications of scores assigned to the test takers by different raters indicated a few important discrepancies between the rater scores. Table 23 and Table 24 present the cross-classification of rating frequencies for two different pairs of raters based on the scores they awarded to two different test takers. Raters 2 and 3 agreed on 1 occasion only out of 12 occasions, and substantial discrepancies were observed on certain dimensions. For example, while Rater 2 assigned 11 on Grammatical

Range and Accuracy to Anna in Task 2, Rater 3 provided a score of 7. In 7 out of 12 ratings, the two raters differed by 2 or more points. Similar findings were observed for a few of the other test takers as well. For example, Raters 1 and 2 differed a lot in terms of the scores they assigned to Jonathan. They agreed on 2 out of 12 cases, and a few of the discrepancies were especially noteworthy. For example, while Rater 1 gave a score of 5 to Jonathan on Lexical Resource in Task 2, the rating given by Rater 2 on the same criteria was only 2. This finding suggests that at least two raters need to be employed even if for the benefit of the doubt.

Table 23. Cross-Classification of Rating Frequencies for Raters 2 and 3 for Anna.

	Rater 3 (score assigned)					
Rater 2 (score assigned)	7	8	9	10	11	Row total
7						0
8						0
9			1			1
10		4	3			7
11	1	1	1	1		4
Column Total	0	5	5	1	1	12

Table 24. Cross-Classification of Rating Frequencies for Raters 1 and 2 for Jonathan.

	Rater 2 (score assigned)				
Rater 1 (score assigned)	2	3	4	5	Row total
2					0
3					0
4	1	4	2		7
5	1	1	3		5
Column Total	2	5	5	0	12

It should be noted that any conclusions drawn from this data must be tentative given the small dataset. The generalizability analyses indicated that the analytic scores would be dependable enough for both low-stakes practice settings and operational use. The D studies indicated that the phi coefficients of the analytic scores were fairly high even for one task with a single rating (.963 –.976) and they improved with double ratings. The improvement was leveled off with each additional task and rater,

which indicated that the test may not need to include six tasks or more than one rater to produce dependable results.

G studies on the scores showed that discrepancies in rater leniency (or severity) accounted for a negligible amount of error (it explained 1.1% of the variance in Fluency scores in Form 1 at its highest). This may be attributed to rater monitoring, as the raters were asked to rate a second time the test takers for whom the scores they assigned showed substantial discrepancies in the first ratings. The raters were also consistent in their rank ordering of test takers, the person-by-rater interaction explaining 0.9%, 1.8%, 1.4%, and 0.7% (in Form 1) and 0.7%, 0.1%, 0.0%, and 0.0% (in Form 2) of the variance of Fluency, Grammatical Range and Accuracy, Lexical Resource, and Coherence/Interaction scores respectively. These results indicate that the raters showed very few discrepancies in the absolute scores they assigned to the test takers.

The variances associated with task main effects and person-by-task interaction were also very low. Even though it has been noted in language assessment literature that some tasks may pose more difficulties for particular sub-groups of test takers (Dunbar, Koretz, & Hoover, 1991), varying results have been reported as to the issue of what makes a task more difficult than another for a certain group of test takers and whether differences in tasks characteristics are reflected in test scores (Bachman, Lynch & Mason, 1995; Elder et al., 2002; Fulcher, 1996b). In the current test, although the tasks were expected to be of different levels of difficulty, this was not reflected in the scores. Lack of score sensitivity to variations in task conditions has been noted in other research studies as well (Fulcher, 1996b; Fulcher & Reiter, 2003; Bachman, Lynch, & Mason, 1995; Lumley & O'Sullivan, 2005). In other words, similar task difficulty levels were observed for different types

of tasks (role play and individual long turn in Bachman, Lynch, & Mason, 1995; interview and group discussion in Fulcher, 1996b). Moreover, variations in task conditions such as social power and degree of imposition (Fulcher & Reiter, 2003) or task familiarity (Lumley & O'Sullivan, 2005) did not produce significantly different scores. In addition, Iwashita et al. (2001) found that differences in assumed degree of cognitive demand based on Skehan's (1998) framework did not predict discourse or score variation. Reviewing a number of studies that have investigated task difficulty, Fulcher and Reiter (2003) write, "research has consistently shown that it requires gross changes in task type to generate significant differences in difficulty from one task to another, and even then the task accounts for little score variance" (p. 326). The findings of the present study also suggest that variations in tasks may not translate into changes in task scores.

Xi and Mollaun (2006) attribute the differences in findings regarding the extent of the person-by-task interaction to the task characteristics and the scoring criteria. They argue,

> If an assessment uses tasks that are not very differentiated in task types and in the ways tasks are contextualized and uses scoring criteria that are more driven by components that are relatively stable across tasks, it is less likely to see variation in performance across tasks. (p. 37-38)

They point out that very contextualized tasks and task-specific scoring criteria are likely to result in large person-by-task interaction. The little task main effect and person-by-task interaction found in the current analysis may be attributed to various reasons. First of all, it may be because of the scoring criteria used. The same scoring criteria were used to rate performances on all tasks (except, of course, that the Coherence dimension was replaced with Interaction in the two tasks that require the test takers to interact with each other). The rating scale does not contain any features that are task-specific; on the contrary, effort was invested in excluding any criteria

that may apply to one task but not another. Fulcher (1996b) proposes that rating scales that do not make references to specific task types, task conditions, or tasks generate scores with most variance accounted for by learner ability. A second possibility is that the tasks were of similar difficulty for this particular group of test takers who participated in the current test, but future administrations of the tasks to different groups may produce more person-by-task variation. A further reason for the little variance attributed to the task main effect and person-by-task interaction may be that the raters tended to assign similar scores to the test takers across tasks based on their overall impression about them on a particular task or the overall test. Teng (2007), for example, found that different scoring methods could show different levels of sensitivity to different task types. She found no difference among different task scores with holistic scoring, while analytic scores produced significant task main effects for complexity and fluency, but not for accuracy. In the present study, in order to minimize possible halo effects, the raters were asked to rate the performances of all test takers on the first task and then continue with the second task and the third task, instead of rating one test taker on all three tasks and continuing with the next test taker. However, if the raters still assigned similar scores across different tasks based on their holistic judgments of the test taker performances, this may be one of the reasons for lack of task main effects.

Overall, the analyses showed that the test taker performances were similar across the tasks based on the scoring criteria used in the test (see also the exemplar performances in Appendix F). It should also be noted here that these findings are consistent with current theoretical models of communicative competence, which, while recognizing the possible influence of context and local components, assert that communicative competence is stable to some extent (Chalhoub-Deville, 2003).

It has been pointed out earlier that G theory addresses the statistical question of the consistency of test taker performances across various tasks and task types. It provides information as to whether the scores obtained from a combination of different tasks can be generalized to the universe of tasks which are similar to those included in the assessment. However, Xi and Mollaun (2006) point out:

> G theory can by no means provide evidence for establishing the link between performance on a sample of tasks ("observed score") and expected performance in the target domain ("target score"), unless there is ample evidence to support that the universe of generalization and the target domain are similar. (p. 39)

From the beginning of the test development process, the issue of validity was the overarching consideration in the current test. The tasks were selected to be representative of the target language use domain. The previous discussions on the validation of the current test have addressed similar concerns. However, another important implication to be drawn from Xi and Mollaun's (2006) argument is that even though a high phi coefficient in the D studies was obtained for one task and one rater condition, because one task would by no means provide enough evidence to argue that the universe of generalization and the target domain are similar, using only one task could not be justified. Therefore, despite the statistical results obtained, several tasks need to be used in the final version of the test. Moreover, as was argued in the previous chapters, the tasks are aimed at eliciting various discourse types and tapping into different language resources even though this difference was not reflected in the scores assigned to the test taker performances.

With regards to the raters, very high inter-rater correlations were found for each of the five dimensions on all of the six tasks. This means that the rank ordering of the test takers did not differ much, and they were similarly rank ordered across all dimension-by-task combinations. The G studies pointed to little rater main effects,

person-by-rater interactions, and task-by-rater interactions. However, because of the limitations associated with rater consensus indices in CTT, and because of the small sample size (12 test takers in each form of the test), which influences results obtained from G and D studies, interpretations based only on these statistics could have been misguided. Investigation of the average agreement rates between the four raters on dimension-by-task combinations revealed a few discrepancies between the raters. Overall, the nonadjacent agreement rates were relatively high and the discrepancies between ratings were more substantial for a few of the dimension-by-task combinations in comparison to others. These discrepancies, however, were produced mainly by two rater pairs out of six rater-rater pairings. In general, the results obtained from the analysis of agreement indicated large agreement rates among the four raters. To better understand the extent of the discrepancies observed in the agreement rate analysis, the scores assigned to each individual test taker were separately analyzed. The cross-classifications of rating frequencies revealed that despite the very high inter-rater correlations, large agreement rates among the raters, and the D study results which produced high phi coefficients for one task-one rater condition, using one rater could not be justified. This is because the summary statistics do not reveal differences at individual level. For example, it was observed that Rater 2 and Rater 3 assigned distinct scores (that differed by 2 or more) in 7 out of 12 dimension-by-task combinations to one of the test takers. If that particular test taker was rated by Rater 2, her scores would be substantially higher than when she would be rated by Rater 3. Similar observations were made for a few of the other test takers, as well. These analyses suggested that at least two raters need to be used for fairness and validity considerations.

Research question 5: To what extent are the test takers' scores related to the scoring dimensions?

4.6.3 Consistency of scores across the scoring dimensions

Table 25 compares the correlations among the four analytic scores by task. As is shown in the table, the correlations among the analytic scores were all very high for all of the six tasks. The correlations were lowest in Task 4 in comparison to the others for all six combinations. However, even in this task the correlations were significant and substantially high.

Table 25. Correlations among the Analytic Scores.

	F vs. A	F vs. L	F vs. C/I	A vs. L	A vs. C/I	L vs C/I
Task 1	.980**	.989**	.990**	.986**	.991**	.990**
Task 2	.986**	.980**	.993**	.986**	.992**	.986**
Task 3	.985**	.985**	.992**	.987**	.983**	.981**
Task 4	.940**	.937**	.967**	.969**	.978**	.980**
Task 5	.970**	.966**	.975**	.979**	.985**	.984**
Task 6	.987**	.983**	.985**	.986**	.992**	.985**

Note. F = fluency, A = grammatical range and accuracy, L = lexical resource, C/I = coherence/interaction.
* $p < .05$, ** $p < .001$.

Although the correlations among the analytic scores on individual tasks were very high, varied profiles are still likely to exist (Xi & Mollaun, 2006). Table 26 shows observed within-person profile variability. It demonstrates the proportion of any two analytic scores differing by 1 or more and by 2 or more by rater and by task. Overall, for over half of the cases (overall by task: 58.5%), any two analytic scores differed by 1 or more. About 9% of the cases had two analytic scores differing by 2 or more. Considering that score differences by 2 move a test taker between levels while score differences by 1 may not (See the rating scale in Appendix C), it can be concluded the test takers showed varied profiles on 9% of the cases on some of the tasks based on the discrepant scores assigned to them on different dimensions.

Table 26. Proportion of Any Two Analytic Scores Differing by 1 or More and by 2 or More.

		1 or 1+	2 or 2+
Task 1	Rater 1	56.9	9.7
	Rater 2	47.2	0.0
	Rater 3	73.6	12.5
	Rater 4	56.9	6.9
	Overall by rater	58.6	7.2
Task 2	Rater 1	54.1	12.5
	Rater 2	43.0	5.6
	Rater 3	65.2	13.8
	Rater 4	54.1	13.8
	Overall by rater	54.1	11.4
Task 3	Rater 1	59.7	12.5
	Rater 2	43.0	4.2
	Rater 3	72.2	16.7
	Rater 4	56.9	9.7
	Overall by rater	58.0	10.8
Task 4	Rater 1	69.4	19.4
	Rater 2	65.2	6.9
	Rater 3	73.6	11.1
	Rater 4	69.4	12.5
	Overall by rater	69.4	12.4
Task 5	Rater 1	56.9	9.7
	Rater 2	50.0	2.8
	Rater 3	51.4	2.8
	Rater 4	65.3	15.3
	Overall by rater	55.9	7.7
Task 6	Rater 1	58.3	4.2
	Rater 2	43.1	1.4
	Rater 3	62.5	5.6
	Rater 4	56.9	2.8
	Overall by rater	55.2	3.5
Overall by task	Overall by rater	58.5	8.8

Based on the correlations among the analytic scores, it can be concluded that the four dimensions are not really distinct from each other as they were reflected in the performance of the particular group of test takers who took the current test. During the rating scale development process, the issue of distinction between the dimensions was addressed through a careful analysis of the rubrics such that the descriptors would not overlap. The distinction between the dimensions used in the rating scale for this test is also made in CEFR and by various test-providers such as Cambridge ESOL. Xi and Mollaun (2006, p. 42-43) argue that very high correlations among the dimensions may be attributed to a number of factors: (a) the correlation among the constructs assessed by each dimension, (b) the dimensions are conceptually distinct,

however the raters could not consistently interpret the descriptors for each dimension (this may be because they could not distinguish among them or because there was overlap among the descriptors which was not captured during the rubric development), or (c) the dimensions are distinct, the rubrics clearly define each dimension and address the conceptual distinction between the dimensions, and the raters were able to differentiate between the dimensions; however, the high correlations among their scores may be attributed to the similarities in the test takers' language learning background, educational background, or exposure to Turkish, to name a few of the possible factors. The scale dimensions, therefore, appear to be closely related. Similar results were obtained for IELTS subscales as well (Jones, 2000, cited in Taylor & Galaczi, 2011). The analysis into the differences between the scores assigned to different dimensions for individual test taker performances, which is presented in Table 26, showed that around 9% of the any two analytic scores differed by 2 or more. This suggests that the test takers' performances on different analytic dimensions were not always in exact agreement. Profile variability analysis pointed to some varied profiles both at individual task levels and at the overall proportion of variability. It should, however, be noted once again that the dimensions were not found to be substantially distinct for this group of test takers; a test taker who performed well in one dimension tended to have a high score for other dimensions as well, while a test taker with lower performance tended to get lower scores for all dimensions. As was mentioned above, this may be attributed either to the sample of test takers, to the analytic scale used, the quality of the ratings, or to a combined effect of these factors (Xi & Mollaun, 2006).

CHAPTER 5

CONCLUSION

5.1 Summary and conclusion

The test in the current study is being developed for the purpose of assessing the speaking proficiency of learners of Turkish as an L2 and is intended to be used at Boğaziçi University to determine the L2 Turkish speaking proficiency levels of incoming exchange and Erasmus students. The construct of speaking assessed in the test was defined based on the communication required in and around the academic setting where the incoming students would spend their time while in Turkey. The test taker characteristics were defined based on a Learner Profile Form, which the test takers filled in before the test development, so that the test could take into consideration test taker characteristics in test development and revision.

The study focused on cognitive demands, contextual features, and scoring validity (reliability) of the test tasks. The discussion of cognitive task demands drew upon Field's (2011) adaptation of the cognitive processing model devised by Levelt (1989). The test tasks were evaluated in terms of the cognitive task demands required by conceptualization, grammatical encoding, phonological encoding, and self-monitoring, as well as time pressure and reciprocity conditions. Since the current test covers a range of proficiency levels, an important consideration was whether the tasks reflected this range in terms of the cognitive demands they impose on the test takers. The tasks were argued to differ in terms of the conceptualization they require from the test takers and the support provided to the test takers. The conceptualization of the broader discourse in test taker performances is captured by the Interaction and Coherence criteria in the rating scale as well as by the requirements of different task types (such as requiring awareness of the ongoing topic or following what the

interlocutor has said). In terms of grammatical encoding, the test tasks were argued to require from the test takers a wide range of functions to be performed, from presentation of simple facts to giving opinions, negotiating, and speculating. The cognitive demands of phonological encoding are captured in the Fluency scale. At lower levels, dysfluency is attributed to language processing in the Fluency scale descriptors, while at higher levels pauses are explained by content planning. Pronunciation, which relates to phonetic encoding and articulation, was not included as an assessment dimension in the test, since based on the analysis of test taker performances, consultation with experts, and an investigation into the literature on the pronunciation of Turkish by learners, it was decided that pronunciation did not pose much difficulty for even learners at low levels of proficiency in terms of intelligibility. Self-monitoring is partially captured in the Fluency rubrics, which make references to communication repair, reformulations, and self-corrections. However, the current test did not directly assess aspects of self-monitoring, since, as suggested by Field (2011), "self-monitoring and self-repair are aspects of speaker performance which are difficult to capture in the form of test specifications" (p. 96). With regard to time pressure, the current test allows planning time in only one of the tasks, which is considered to be conceptually more difficult and thus to place heavier conceptualization demands on the test takers. The other tasks are interactive tasks which replicate real life speaking events, which take place under time pressure and require spontaneous production/interaction. Lastly, the test uses various interaction formats and creates different reciprocity conditions. The possible effects of power difference between the interviewer and the test takers in test performance is mitigated by employing test taker-test taker format. Moreover, the roles of the

interlocutor (whether this is the examiner or a test taker pair) and the test taker change from one task to the other.

The discussion of contextual task characteristics were built upon the aspects of context validity as outlined by Weir (2005) and operationalized by Galaczi and ffrench (2011). In this regard, the tasks used in the current test were critically analyzed in terms of the variables related to the task setting and task demands. It was argued that the test includes various response formats (interview, test taker-test taker interaction, individual long run), and thus elicits a wide range of speech functions, provides varied production and interaction opportunities, and assigns different roles to the test takers. The test includes different task purposes (description, narration, commenting, explanation, justification, etc.), and task purpose is directly communicated to the test takers in the task instructions. All tasks and all assessment dimensions contribute equally to the overall score assigned. If one aspect of speaking ability is observed to be more important than others in the Turkish Language Program at Boğaziçi University, certain tasks and criteria may be weighted more in future administrations of the test. This requires a close investigation of the program aims, requirements, and implementation in terms of whether one aspect of speaking ability is prioritized over others; however, the current study did not feature this particular investigation.

The direct testing approach adapted in the test enables interaction, in addition to production, so communication is co-constructed by the test takers; and different input channels (aural, written, visual) are utilized in the test. The input length in the instructions were revised many times so that instructions are brief and easy to follow. However, in three of the tasks, the input length had to be longer. In these cases, the sentences were kept grammatically and lexically simple. In terms of the nature of

information, the test employs tasks that vary on the personal/non-personal and abstract/concrete continuum. One task was found to be problematic with regard to topic familiarity. Although the analysis of the scores did not indicate bias towards a certain sub-group of test takers, qualitative analysis showed that one of the two pictures in Task 2 might be biased against female test takers, and the other may be distressing to some test takers. This pointed to the need to change these two pictures in future forms of the test. Lexical and structural resources and their relation to proficiency are captured by the Lexical Resource, and Grammatical Range and Accuracy criteria in the rating scale. Interlocutor and administration variables were not investigated in the current study.

The scoring validity of the test was analyzed based on the parameters proposed by Taylor and Galaczi (2011) and using generalizability theory. The descriptions of the levels in the rating scale were adapted from the CEFR level descriptions. Aspects of rater variability were addressed through rater training and rater monitoring. Still, a few discrepancies were noted between the scores assigned by the raters. To analyze rater effects, the dependability of analytic scores, and the effect of the number of tasks and raters on score dependability, generalizability and decision studies were carried out. It is important to note once again that given the small data set, the conclusions drawn are tentative. First of all, high score dependability was obtained for all of the six tasks used. The D studies indicated high score dependability even for one task and one rating condition. However, problems with this interpretation were noted. The one task condition could not be justified, not least because the tasks tap into different language resources, and different discourse and interaction types. Moreover, despite the very high inter-rater correlations, an investigation into the scores assigned to the individual test takers by different raters

through the cross-classifications of the ratings showed that there were important discrepancies in a few cases, and this suggested that at least two raters need to be used for fairness and validity considerations. This concern can be justified also because these discrepancies were observed after rater training and rater monitoring. Besides, the analyses showed that the scoring dimensions did not produce different scores from each other, but they were not in exact agreement, either. It was argued that this finding could be attributed to the sample of test takers, the nature of the analytic rubric used, the quality of analytic ratings, or a combined effect of these.

The test tasks used in the test were expected to be of different difficulty levels, yet this was not reflected in the test scores. This was argued to be attributable to various reasons. It may be due to the scoring criteria used, which do not contain task-specific features. It may also be that the tasks were of similar difficulty for the specific group of test takers who participated in the study, and may produce different scores with another group. A further reason may be halo effect, despite the attempts made to minimize such an effect. However, it should be noted that these results are consistent with the findings of previous studies, which suggest that variations in task conditions may not necessarily translate into changes in scores.

It was found that the person-by-rater interaction was relatively higher in the two picture description tasks in comparison to the other four tasks. However, even in these two tasks, the interaction was very small. These tasks may produce higher person-by-rater interaction with a bigger number of test takers; however, the tasks all seem to be fulfilling their purpose. The statistical analyses did not provide much information about which task(s) to keep, which one(s) to revise, or which one(s) to abandon. However, as is pointed out earlier, the qualitative analyses pointed to

problems with one of the picture tasks, Task 2. Different pictures need to be used in this task in future forms of the test.

5.2 Limitations

As has been noted a few times earlier, one of the limitations of the current study is the limited sample size. The conclusions drawn from the analyses were, therefore, tentative. This aspect must have impacted the results obtained with regard to the tasks, raters, and rating dimensions. For example, the high dependability scores obtained and the little variance associated with rater main effects and rater-by-person interaction were not unexpected given the small number of test takers taking each form of the test (12 persons in Task 1, and 12 in Task 2), and the heterogeneity of the test takers in terms of proficiency levels. Because of practicality reasons, the test had to be divided into two forms, and each form was taken by one half of the test takers. As a result, the number of participants taking each task was 12, instead of 24.

The rating scale used in the study was adapted from the CEFR level descriptors. With a bigger performance data, the rubric could have been based on empirical findings from the L2 Turkish performance of test takers with varying proficiency levels, and the CEFR scales could be 'fine-tuned' to reflect the context of the current test. However, because there were only a few test takers at each level, their data could not be used to make generalizations about performance 'representative' of a proficiency level.

Lastly, the current study does not adequately address practicality considerations. A test that includes all of the six tasks may be difficult to administer in practical settings. Therefore, as is noted above, another study should be carried out, focusing again on the validity and dependability of different numbers and combinations of the tasks with more participants. Moreover, G and D studies may be

conducted with and without rater monitoring by the researcher, since an analysis into the initial rater scores are likely to produce different results in generalizability analysis.

However, with the substantial support from theory and careful design employed in its development, this test is a promising test to be used for the purpose of assessing the speaking proficiency levels of incoming exchange/Erasmus students at Boğaziçi University. Taylor (2011) highlights the importance of test providers' being "transparent and accountable in what they do, especially in terms of the standards to which they adhere, the quality and validity claims of their products, and the provision of theoretical and empirical evidence in support of these claims" (p. 5), and the current study does just that. The a priori validation arguments regarding the test provide substantial evidential basis for the quality of the test. Moreover, the study describes in detail the theoretical basis for the test and provides a posteriori evidence for the scoring validity of the test.

5.3 Suggestions for further research

As has been pointed out, this was a small scale study with a small number of participants. A further step for the development of current test such that it is ready for operationalized use is to conduct revisions on the problem areas that have been detected in the current study, such as replacing the pictures in Task 2 with another set of pictures, and to pilot the novel form of the test with a larger sample.

The rating scale used in the current study was drawn from the CEFR level descriptors. A study with a larger sample may provide a larger amount of learner language data on which further revisions and 'fine-tuning' of the rating scale can be based. The scale used in the current study did not include pronunciation as an assessment dimension. However, this does not suggest that pronunciation is not an

important aspect of communicative language ability. Further research is needed that investigate the development of pronunciation skills of learners of Turkish as a second language and more substantial aspects of Turkish pronunciation such that scoring rubrics for pronunciation can be based on these empirical data.

Criterion-related validity and consequential validity issues were not investigated in the current study. Both concurrent and predictive validity studies can be carried out for stronger validity claims. For example, test taker performances can be followed when the test takers are administered to the TKF classes at Boğaziçi University. Investigating the positive or negative consequences of the test for the incoming students and the Turkish Language Program at Boğaziçi University may be the focus of other research.

APPENDIX A

TEST TASKS

A1 The tasks used in the current study

The final version of the tasks used in the second pilot examination are presented below. Then, the initial tasks which were either excluded from the test or were revised to be included in the second pilot examination will be provided.

TEST FORM 1

The Form 1 included Task 1, Task 2, and Task 3. Note that the numbering of the tasks does not indicate a difficulty cline.

TASK 1: INTERVIEW

Bu bölüm öğrenci ile karşılıklı konuşma şeklinde olacak. Bu bölüm için 7 dakika ayırın. Bireye A, B ve C bölümlerinin her birinden sorular yönlendirin. Öğrenci yavaş ise, en son bölümden de soru sormak için diğer bölümlerden bazı soruları atlayabilirsiniz. Örneğin C bölümünde kitaplarla ilgili soruyu sorup filmlerle ilgili olanı geçebilirsiniz.

Öğrenciye yönergeler:
"Hoş geldiniz. Sınavın bu bölümü soru-cevap şeklinde olacak. Ben size bazı sorular soracağım. Bu sorulara cevap vermenizi istiyorum."

A)

- Merhaba, hoş geldiniz? İsminiz nedir?
- Boş zamanlarınızda neler yaparsınız?
- Hangi alanda eğitim görüyorsunuz? (Ne okuyorsunuz?)
- Bu alanı neden tercih ettiniz?
- Alanınızda en çok hangi konular ilginizi çekiyor? (Söylenilen konulardan birini seçip o konuyu kısaca açıklamasını isteyin)
- Gelecekte nasıl bir işte çalışmak istiyorsunuz? Neden?

B)

- Günde kaç saatinizi internette harcıyorsunuz?
- İnterneti daha çok ne için kullanıyorsunuz? (İnternette hangi konulara ilgi duyuyorsunuz?)

C)

- Kitap okumayı sever misiniz? Hangi kitapları okursunuz? (Favori kitabınız hangisi? Kitapta neler oluyor?)
- Hangi film türlerinden hoşlanıyorsunuz? (Favori filminiz hangisi? Bu filmde neler oluyor?)
- Film ve kitaplara sansür uygulanması hakkında ne düşünüyorsunuz?
- Sizce hangi durumlarda sansür uygulanmalıdır?

D)

- Ülkenizin eğitim sisteminde yaşanan en büyük sorunlar nelerdir? Bu konuda diğer ülkeler ile kendi ülkenizi karşılaştırabilir misiniz?
- Dünyanın birçok ülkesinde nüfus yaşlanıyor. Sizin ülkenizde durum nedir? Yaşlı nüfusun artmasının sizce ne gibi sonuçları olabilir? Bu nasıl sorunlar yaratabilir?

TASK 2: EVENTS

Bu bölümde öğrencilere resimleri vererek aşağıdaki soruları yönlendirin. Bu bölüm için 4 dakika ayırın.

Bu iki resmi inceleyin. Size 4 soru soracağım. Her soru için 1 dakikanız var.

1) *Lütfen resimlerde gördüklerinizi anlatın.*

2) *Sizce bu insanlar nasıl hissediyor?*

3) *Sizce bu resimler çekilmeden önce neler oldu?*

4) *Sizce bundan sonra ne olacak?*

1



2



TASK 3: POSTER

Bu bölümde öğrenciler birbirleri ile konuşacaklar. Aşağıdaki yönergeleri verin. Anlamazlarsa ikinci kez tekrar edin.

Görüşmenin bu bölümünde birbirinizle konuşmanızı istiyorum. Şimdi resimlere bakın.

Bir üniversite yeni ders yılı için üniversiteyi temsil edecek bir poster hazırlıyor. Bunun için bir araştırma yapıyor. Bu posterde kullanmak için bazı alternatif fotoğraflar seçiyorlar. Sizce fotoğrafların her birinde üniversitenin hangi yönü gösteriliyor? Sizce bu beş resimden hangisi poster için en iyi seçenek olur? Aranızda tartışın ve bir resim seçin.

Bunun için 5 dakikanız var. Lütfen başlayın.

1



2



3



4



5



TEST FORM 2

Form 2 included Task 4, Task 5, and Task 6.

TASK 4: HOLIDAY

Bu bölümde, size dört resim vereceğim. Üç soru soracağım. Her soru için bir dakikanız var. (Resimler verilir).

1) *Lütfen, bu resimlerde gördüklerinizi anlatın. Resimlerde ne görüyorsunuz?*

2) *Sizce bu insanlar neden bu tatil yerlerini seçti?*

3) *Siz bu tatillerden hangisini seçerdiniz? Neden?*









TASK 5: EXTENDED TALK

Öğrenciye konuları yazılı olarak veriniz. Seçtiği konuyu sesli olarak bir kez de siz okuyunuz.

Bir seminerde 3 dakikalık bir konuşma yapacaksınız. Size üç konu vereceğim. Birini seçin ve bu konu hakkında 3 dakika konuşun. Hazırlık için 1 dakikanız var.

Şimdi lütfen konulardan birini seçin.

(Seçilen konuyu öğrenci için bir kez de siz okuyunuz)

Şimdi bir dakika hazırlanabilirsiniz. Not alabilirsiniz. Sonra 3 dakika konuşacaksınız.

1. "Küresel ısınmayla birlikte Dünya bir kuraklık tehlikesiyle karşı karşıya. Sizce bunu önlemek için ne yapılmalı? Bireysel düzeyde ve hükümetler düzeyinde neler yapılabilir?"

2. "Sosyoekonomik durum insanların yaşam tarzlarını nasıl etkiler? Örnekler vererek açıklayınız."

3. "Ülkenizde göçmenler var mı? Daha çok hangi ülkelerden geliyorlar? Sizce hem kendi ülkenizde hem de diğer dünya ülkelerinde göçmenlerin en büyük problemleri nelerdir?"

TASK 6: SCHEDULE

Size birer ders programı vereceğim. İkinizde de farklı bir ders programı var. Birlikte İKİ adet seçmeli ders almak istiyorsunuz. Programların altındaki seçmeli derslere bakın ve ikinizin programına da uygun iki seçmeli ders seçin.

Bu bölüm için 5 dakikanız var.

Program 1

	10:00-10:50	11:00-11:50	12:00-12:50	13:00-13:50	14:00-14:50	15:00-15:50	16:00-16:50	17:00-17:50
Pazartesi				Sosyolojiye Giriş			Modern Tarih Okumaları	
Salı	Sosyal Bilimlerde Araştırma Teknikleri					Fotoğrafçılık		
Çarşamba	Kültür ve Toplum		Ticari Gazeticilik					
Perşembe	Sosyolojiye Giriş					Modern Tarih Okumaları		
Cuma		Kültür ve Toplum			Yabancı Dil			

SEÇMELİ DERSLER:

Pazartesi, 12:00-13:50 _ Başlangıçtan Bugüne Avrupa Sineması

Pazartesi, 12:00-13:50 _ Shakespeare Okumaları

Salı, 12:00-15:50 _ Sinema Tarihi

Salı, 13:00- 14:50 _ Cumhuriyetçilik, Özgürlük ve Demokrasi Okumaları

Çarşamba, 15:00- 16:50 _ Felsefe Tarihi

Çarşamba, 16:00- 17:50 _ Toplum Psikolojisi

Çarşamba, 15:00- 17:50 _ Yaratıcı Yazarlık

Perşembe, 12:00-13:50 _ Hızlı Okuma Teknikleri

Perşembe, 13:00-14:50 _ Etkili İletişim Teknikleri

Perşembe 13:00-14:50_ Modern Türk Edebiyatı

Cuma, 10:00-11:50 _ Resim

Cuma, 10:00-10:50 / 11:00-11:50_ Evrenin Oluşumu ve İnsan

Program 2

	10:00-10:50	11:00-11:50	12:00-12:50	13:00-13:50	14:00-14:50	15:00-15:50	16:00-16:50	17:00-17:50
Pazartesi					Müzik Kulübü Toplantısı			
Salı	Yabancı Dil				Psikolojiye Giriş			
Çarşamba			Yabancı Dil					
Perşembe	Klasik Müziğin Tarihi					Beyin ve Bilişim		
Cuma			Çocuk Gelişimi					

SEÇMELİ DERSLER:

Pazartesi, 12:00-13:50 _ Başlangıçtan Bugüne Avrupa Sineması

Pazartesi, 12:00-13:50 _ Shakespeare Okumaları

Salı, 12:00-15:50 _ Sinema Tarihi

Salı, 13:00- 14:50 _ Cumhuriyetçilik, Özgürlük ve Demokrasi Okumaları

Çarşamba, 15:00- 16:50 _ Felsefe Tarihi

Çarşamba, 16:00- 17:50 _ Toplum Psikolojisi

Çarşamba, 15:00- 17:50 _ Yaratıcı Yazarlık

Perşembe, 12:00-13:50 _ Hızlı Okuma Teknikleri

Perşembe, 13:00-14:50 _ Etkili İletişim Teknikleri

Perşembe 13:00-14:50_ Modern Türk Edebiyatı

Cuma, 10:00-11:50 _ Resim

Cuma, 10:00-10:50 / 11:00-11:50_ Evrenin Oluşumu ve İnsan

A2 The initial tasks

The following tasks were developed in the initial phase of the test development. These tasks were either excluded from the test or were revised to be included in the second pilot examination based on expert feedback and the results obtained from the first pilot examination.

All of these tasks were developed based on the level descriptions drawn from the CEFR. The table below shows the topics drawn from the CEFR and the task types that could be used in relation to these topics.

	TOPICS DRAWN FROM THE CEFR	POSSIBLE TASK TYPES
A2	Familiar and routine matters related to work and free time Very short social exchanges Structured situations and short conversations Familiar topics in predictable everyday situations *people, living or working conditions, daily routines, likes/dislikes, etc.	- a short interview (introducing himself/herself; questions about likes/dislikes; questions about school and courses) - picture description - talking about daily routines
B1	Travelling Topics that are familiar, of personal interest or pertinent to everyday life (family, hobbies, work, travel, and current events). His/her interests and professional field Explaining why something is important More abstract, cultural topics such as films, books, music, etc. One of a variety of subjects within his/her field	- comparing and contrasting two kinds of holiday accommodations/ holiday activities/ ways of traveling, etc. (pictures?) - choosing among different genres of movies/books/music and explaining why s/he prefers them - The interviewer wants to buy a book/film (music) CD for a friend of him/hers and asks the examinee to offer him/her a few and explain why - Explaining why s/he preferred his/her field of study; why s/he is interested in it; what are some particular topics that interest him/her most in this field; briefly explain those topics - information gap activities (paired)

B2	The personal significance of events and experiences Account for and sustain views clearly by providing relevant explanations and arguments General, academic, vocational or leisure topics, marking clearly the relationship between ideas A wide range of subjects related to his/her field of interest, expanding and supporting ideas with subsidiary points and relevant examples Systematically developed descriptions and presentations—significant points, relevant supporting detail	- University poster (paired discussion) - reading a graph and talking about it - talking about a pre-planned topic (smart phones; ways to create a peaceful environment; what makes a university a good one (better than another) etc. - the examinee is presented with a problem and is asked to suggest solutions to it, explaining why these are good solutions (For instance: a university has only enough space to be used as either a cafeteria (dining hall), or a fitness center or to extend its library. There is need for all of them; but one needs to be chosen. Which one would you choose and why?)
C1	Only a conceptually difficult subject can hinder a natural, smooth flow of language Detailed descriptions and presentations on complex subjects, integrating sub-themes, developing particular points and rounding off with an appropriate conclusion	- a presentation on a pre-planned topic (the topic needs to be a complex one, preferably an academic one)

Initial Task 1:

Bu bölüm soru-cevap şeklinde geçecektir.

- Merhaba, hoş geldiniz? İsminiz nedir?

- En büyük hobiniz nedir? (Bunu neden seviyorsunuz?)

- Partilerden hoşlanır mısınız? (Gittiğiniz bir partiyi kısaca anlatınız)

- Boş zamanlarınızı kiminle geçiriyorsunuz? (Birlikte ne tür aktiviteler yapıyorsunuz?)

- Son zamanlarda çok güzel zaman geçirdiğiniz bir günü anlatın.

- Haftada kaç saatinizi internette geçiriyorsunuz?

- İnterneti daha çok ne için kullanıyorsunuz? (Bunlardan nasıl faydalanıyorsunuz?)

- İnterneti yeni şeyler öğrenmek için kullanıyor musunuz? Ne tür konulara ilgi duyuyorsunuz?

- Hangi alanda eğitim görüyorsunuz? Bu alana neden ilgi duyuyorsunuz?

- Bu alanda öğrenmek istediğiniz yeni şeyler var mı? Bunlar neler?

- Gelecekte ne tür bir işte çalışmak istiyorsunuz? Neden?

Initial Task 2:







Bu resimlerde farklı insanlar alışveriş yaparken gösteriliyor. Bu insanların alışveriş alışkanlıklarını, tercihlerini anlatınız. Daha sonra fotoğraflardan bu insanlarla ilgili ne tür çıkarımlar yaptığınızı anlatınız.

Initial Task 3:







Bu üç resimde gösterilen insanların kimler olduklarını, içinde bulundukları durumu, duygu durumlarını ve resimde gösterilen olayın öncesinde neler olmuş olabileceğini anlatınız.

Initial Task 4:

Aşağıda çeşitli film türleri listelenmiştir. Bu türlerden hangisini en çok beğendiğinizi belirtiniz. Sonra bu türü neden seçtiğinizi örneklerle açıklayınız.

AKSİYON FİLMLERİ

KOMEDİ FİLMLERİ

FANTASTİK FİLMLER

TARİHİ FİLMLER

KORKU FİLMLERİ

FELSEFİ FİLMLER

POLİTİK FİLMLER

ROMANTİK FİLMLER

BİLİM KURGU FİLMLER

POLİSİYE FİLMLER

ANİMASYON FİLMLERİ

Initial Task 5:





Sınavın bu bölümünde, farklı tatil türlerini karşılaştırmanızı istiyorum. Size iki resim vereceğim. (Resimler verilir). Öncelikle, bu resimlerde gördüklerinizi anlatınız. Sonra, resimlerdeki insanlar hakkında ne tür çıkarımlarda bulunduğunuzu ve son olarak hangi tatil türünü tercih ettiğinizi sebepleriyle açıklayınız.

Initial Task 6:

Bir arkadaşıma hediye olarak bir kitap almak istiyorum. Bu edebi bir kitap, bilimsel bir kitap, sanat kitabı vb. olabilir. Bana bir kitap önermenizi ve bu kitabı anlatmanızı istiyorum.

Initial Task 7:

İkinize de birer ders programı ve seçmeli derslerin olduğu bir liste vereceğim. Birlikte bir seçmeli ders almak istiyorsunuz. Programlarınızı inceleyin. İkinizin programına da uygun olan seçmeli dersleri bulun ve bunlardan birine karar verin.

	10:00-10:50	11:00-11:50	12:00-12:50	13:00-13:50	14:00-14:50	15:00-15:50	16:00-16:50	17:00-17:50
Pazartesi				Sosyolojiye Giriş			Modern Tarih Okumaları	
Salı	Sosyal Bilimlerde Araştırma Teknikleri					Fotoğrafçılık		
Çarşamba			Ticari Gazetecilik					
Perşembe	Sosyolojiye Giriş					Modern Tarih Okumaları		
Cuma		Kültür ve Toplum			Yabancı Dil			

	10:00-10:50	11:00-11:50	12:00-12:50	13:00-13:50	14:00-14:50	15:00-15:50	16:00-16:50	17:00-17:50
Pazartesi		Psikolojiye Giriş			Müzik Kulübü Toplantısı			
Salı	Yabancı Dil					Psikolojiye Giriş		
Çarşamba			Yabancı Dil		Klinik Psikoloji			
Perşembe	Klasik Müziğin Tarihi					Beyin ve Bilişim		
Cuma			Çocuk Gelişimi					

SEÇMELİ DERSLER:

Pazartesi, 11:00-11:50 / 12:00-12:50 _ Başlangıçtan Bugüne Avrupa Sineması

Pazartesi, 11:00-11:50 /12:00-12:50 _ Sinema Tarihi

Salı, 12:00-12:50/ 13:00-13:50 /14:00-15:50 _ Shakespeare Okumaları

Salı, 13:00-13:50 /14:00-14:50 _ Cumhuriyetçilik, Özgürlük ve Demokrasi Okumaları

Çarşamba, 10:00-10:50 /11:00-11:50 _ Tenis

Çarşamba, 10:00-10:50 /11:00-11:50 _ Toplum Psikolojisi

Çarşamba, 15:00-15:50 /16:00-16:50 /17:00-17:50 _ Günlük Hayat ve Birey

Perşembe, 13:00-13:50 /14:00-14:50 _ Yüzme

Perşembe, 12:00-12:50 / 13:00- 13:50 _ Yoga

Cuma, 10:00-10:50 /11:00-11:50 _ Resim

Cuma, 10:00-10:50 / 11:00-11:50_ Evrenin Oluşumu ve İnsan

Initial Task 8:





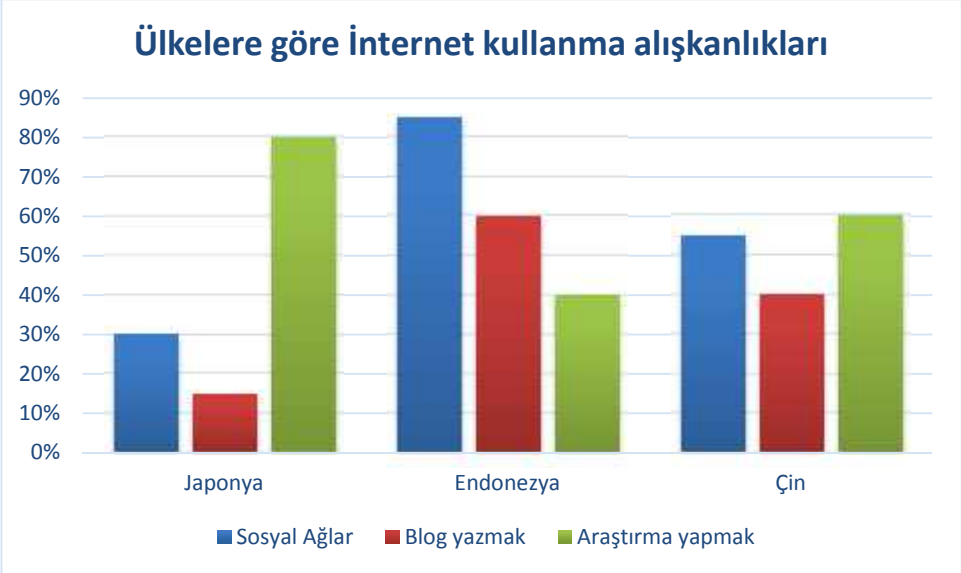



Testin bu bölümünde birbirinizle konuşmanızı istiyorum. Şimdi resimlere bakın.

Bir üniversite yeni ders yılı için poster çalışması yapıyor. Bu posterde kullanılmak üzere bir fotoğraf seçilecek. Resimlerin her birinde üniversitenin hangi yönünün gösterildiğini tartışın. Sonra bu altı resimden hangisinin poster için en iyi seçenek olduğuna karar verin.

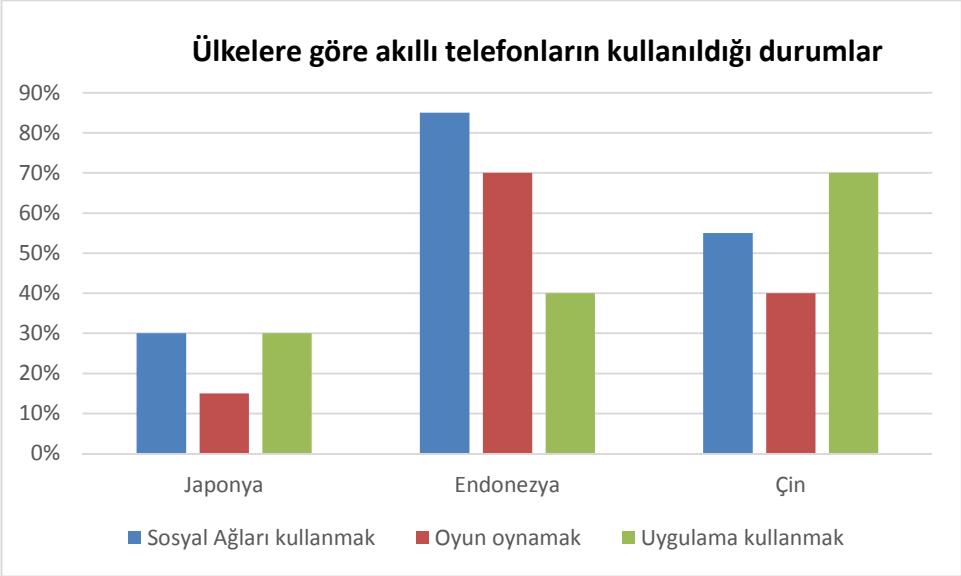
Bunun için 3 dakikanız var. Lütfen başlayın.

Initial Task 9:



Bu tabloda üç ülkenin internet kullanma alışkanlıkları gösterilmektedir. Her bir ülkenin internet kullanma alışkanlıklarını açıkladıktan sonra, ülkeleri birbiri ile karşılaştırınız. Sonrasında kısaca kendi internet kullanma alışkanlıklarınızı anlatınız.

Initial Task 10:



Üç ülkede akıllı telefon kullanıcılarına telefonlarını şu üç durum için kullanıp kullanmadıkları sorulmuştur:

- Sosyal ağları kullanmak,

- Oyun oynamak,

- Uygulama kullanmak.

ve tabloda gösterilen sonuçlar elde edilmiştir. Akıllı telefonların kullanıldığı durumlar açısından bu üç ülkeyi karşılaştırınız.

Initial Task 11:
Akıllı telefonların hayatımızı nasıl etkilediğini anlatınız. Akıllı telefon kullanımının avantajları ve dezavantajları nelerdir?

Initial Task 12:

Sizin için iyi bir üniversitenin özellikleri nelerdir? Bir üniversiteyi diğerlerinden daha iyi ve başarılı yapan özellikleri örnekler vererek anlatınız.

Initial Task 13:

Bir üniversitede öğrenciler için bir yemekhaneye, bir spor salonuna ve kütüphanenin genişletilmesene ihtiyaç vardır. Ancak üniversite bunlardan yalnızca birini gerçekleştirebilecek imkânlara sahiptir. Siz hangisini yapmayı tercih ederdiniz? Sebepleriyle açıklayınız.

Initial Task 14:








Bu bölümde size bazı fotoğraflar vereceğim. Bu fotoğraflar ile ilgili olarak iki dakika konuşmanızı istiyorum. Sonrasında fotoğraflarla ilgili bir soru soracağım ve bunları cevaplamanızı isteyeceğim.
Fotoğraflar akıllı telefonların hayatımızı nasıl etkilediğini gösteriyor. Fotoğraflarda gösterilen değişiklikleri anlatın.
(Ardından aşağıdaki sorulardan biri sorulur.)

- Akıllı telefonların insanların sosyal iletişim becerilerini azalttığı söyleniyor. Sizin bu konudaki fikriniz nedir?
- Akıllı telefon kullanımının bazı sağlık sorunlarına sebep olduğu söyleniyor. Sizin bu konudaki fikriniz nedir?
- Akıllı telefonların getirdiği en önemli yenilik sizce nedir? Neden?

Initial Task 15:










Testin bu bölümünde size bazı resimler vereceğim. Bu resimlerden bir hikaye oluşturmanızı isteyeceğim. Hikayeyi mümkün olduğunca detaylı anlatmaya çalışınız. Bunun için bir dakikanız var.

APPENDIX B

TASK SPECIFICATIONS

One of the most detailed lists of contents for test specifications was developed by Alderson, Clapham and Wall (1995). According to them, the specifications should include: the test's purpose; description of the test takers; test level; definition of construct (theoretical framework for the test); description of suitable language course or textbook; number of sections/papers; time for each section/paper; weighting for each section/paper; target language situation; text-types; text length; language skills to be tested; language elements to be tested; test tasks; test methods; rubrics; criteria for marking; descriptions of typical performance at each level; description of what candidates at each level can do in the real world; sample papers; samples of student performance on task (p. 11-20).

The test's purpose; description of the test takers; test level; definition of the construct; number of sections; weighting for each section; the target language situation; language skills to be tested; language elements to be tested; test tasks; test method; rubrics; criteria for marking have all been provided in the main body of the present text. Below, each task is described briefly, and the task specifications are presented.

Task 1 - Interview

In this task, the test takers are expected to respond to the examiner's questions. This task gives the test takers the opportunity to show their ability to use general social and interactional language and talk about themselves and their interests, experiences, plans for the future, their field of study, opinions on a topic, etc. The task assesses the test takers' ability to take part in spontaneous communication in an everyday setting. The examiner asks the test takers for some information about themselves, then widens the scope of the questions by asking about, e.g. the test takers' leisure activities, hobbies, daily life, experiences, interests, majors. This kind of social exchange is a natural way to begin an interaction before dealing with more specific tasks. The test takers should be discouraged from giving one-word answers in this part; especially when asked about their daily routines, likes/dislikes, or plans for the future. The test takers should be encouraged to extend their answers with reasons and examples. O'Sullivan (2008) lists the advantages of structured interviews as providing greater possibility of all test takers being asked the same questions, and therefore, rendering comparisons more valid; high degree of content and face validity; and high inter- and intra-rater reliability with training. The disadvantages of this task type listed by O'Sullivan are covering a limited range of situations and that examiners may introduce other questions than the predetermined ones. The examiners' way of interacting with the test takers may also produce construct-irrelevant score variance (Brown, 2003; Nakatsuhara, 2008).

Task Specifications	
Task Input/ Prompt	
Task Content (as informed by the CEFR level descriptions)	A2 to C2 CEFR A2: Can ask and answer questions and exchange ideas and information on familiar topics in predictable everyday situations Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters to do with work and free time. - *Boş zamanlarınızda neler yapmaktan hoşlanırsınız?* CEFR B1: Can communicate with some confidence on familiar routine and non-routine matters related to his/her interests and professional field. Can express thoughts on more abstract, cultural topics such as films, books, music, etc. Can enter unprepared into conversation on familiar topics, express personal opinions and exchange information on topics that are familiar, of personal interest or pertinent to everyday life (e.g. family, hobbies, work, travel and current events). - *Hangi alanda eğitim görüyorsunuz? Ne okuyorsunuz? Bu alanı neden tercih ettiniz?* - *Kitap okumayı sever misiniz? Ne tür kitaplar okursunuz? En sevdiğiniz kitabı kısaca anlatınız.* CEFR B2: Can use the language fluently, accurately and effectively on a wide range of general, academic, vocational or leisure topics, marking clearly the relationship between ideas. Can highlight the personal significance of events and experiences, account for and sustain views clearly by providing relevant explanations and arguments. - *Film ve kitaplara sansür uygulanması hakkında ne düşünüyorsunuz?* - *Sizce hangi durumlarda sansür uygulanmalıdır?*
Task Type and Format	Each test taker interacts with the examiner. The examiner asks the test taker questions, using standardized questions.
Task Focus / Target Skills	Giving information on a factual and personal kind. The test taker responds to questions about his/her interests, past experiences and future plans, field of study, and ideas about a topic of debate.
Topic	Personal identification Daily life

	Free time, Entertainment Education
Input	Questions directed orally by the examiner
Channel for task delivery	Face to face only, in real time
Interaction pattern	Dialogue: Test taker/examiner
Task instructions	Spoken (real time)
Type of prompt	Oral only (real time/given by the examiner)
Language level of instructions	Same as the level of the task
Recorded?	The whole task is recorded for rating
Response	
Response type	Interaction (short responses for some questions; extended responses for others)
Timing	7 minutes
Preparation time	0
Interaction required	Interaction with examiner
Discourse type required	Conversation / Question and answer
Rhetorical functions	Reporting events (Factual presentation) Referring to past-present-future Explanation Giving an opinion Giving cause-reason Giving example Comparing and contrasting Expressing modality Speculating Summarization
Register	Neutral (neither informal nor formal)
Domain	Personal, Educational, Academic
Rating Of Task	
Rating method	Analytical method
Rating criteria	Fluency, Range and Accuracy, Lexical Resource, Coherence

Task 2 - Events

In this task, the test takers are given two color photographs to describe. The photographs will present a snapshot from two situations where people are shown to experience different events and feelings. After giving a detailed description of the two pictures, the test takers are asked to hypothesize about (1) what happened/ might have happened before the photograph was taken and (2) what will/ might happen after this moment. The task gives the test takers the opportunity to show their range of vocabulary, and to talk about past and future events, and thus use different tenses (present progressive while describing the events, past tense for talking about what happened before, and future tense to talk about what will happen afterwards). The test takers are required to hypothesize and express possibility/probability. This task, therefore, requires use of a range of functional language. The advantages of such tasks include clarity; eliciting extended sample of connected speech; examining the test takers' ability to use certain grammatical forms; and valid performance comparisons (O'Sullivan, 2008). O 'Sullivan (2008) also lists a number of disadvantages: limited authenticity; requiring little interaction; unclear and ambiguous pictures; and limited score reliability due to differences in test takers' interpretations of the pictures.

Task Specifications	
Task Input/ Prompt	
Task Content (as informed by the CEFR level descriptions)	B1 CEFR B1: Can relate a straightforward narrative or description as a linear sequence of points. Can give detailed accounts of experiences, describing feelings and reactions. Can relate the details of unpredictable occurrences, e.g. an accident. Can describe events, real or imagined. Has a sufficient range of language to describe unpredictable situations. *The test takers are asked to give accounts of two events shown in the pictures, speculating about what might have happened before the moment the picture shows and what may happen afterwards. They are also asked to describe the feelings of the people shown in the pictures.*
Task Type and Format	The examiner asks the test takers questions, using standardized questions. The test taker takes relatively longer turns to answer each question
Task Focus / Target Skills	Describing and commenting on the visual prompt, using appropriate vocabulary, managing discourse in a longer turn, hypothesizing about possible events, and using appropriate grammatical forms. The test taker gives a description of and responds to questions about the visual prompt. S/he hypothesizes about the events preceding and following the moment shown in the visual prompt.
Topic	A particular event (such as an accident, or a sports event) People and feelings Preceding and following events
Input	Pictures Questions directed orally by the examiner
Channel for task delivery	Face to face only, in real time
Interaction pattern	Dialogue: Test taker/examiner Extended responses by the test taker
Task instructions	Spoken (real time)
Type of prompt	Pictorial Oral (real time/given by the examiner)
Language level of instructions	Same as the level of the task
Recorded?	The whole task is recorded for rating
Response	
Response type	Interaction (the examiner asks questions and the test taker gives extended responses)
Timing	4 minutes
Preparation time	0

Interaction required	Interaction with the examiner Extended responses by the test taker
Discourse type required	Question and answer
Rhetorical functions	Description (pictures) Reporting events Referring to past-present-future Expressing modality Speculating
Register	Neutral (neither informal nor formal)
Domain	Personal/Public
Rating Of Task	
Rating method	Analytical method
Rating criteria	Fluency, Range and Accuracy, Lexical Resource, Coherence

Task 3 - Poster

In this task, two test takers are given oral instructions and are provided with visual stimuli (several photographs showing different aspects of a university) to form the basis for a task which they carry out together. They are expected to discuss each visual, expressing and justifying opinions, evaluating and speculating, in order to work towards a negotiated decision towards the end of the task. The instructions make these two parts of the task clear: 'Fist, talk to each other about…', and 'Then decide…' The decision should only be made after the test takers have explored each of the issues as illustrated by the pictures, and they are assessed on their ability to use the language of interaction (negotiation and collaboration). However, they are not penalized if they fail to reach a negotiated decision. There is no right or wrong answer to the task. The task gives the test takers the opportunity to show their range of language and their ability to invite the opinions and ideas of their partner. The test takers are expected to share the interaction in this way and respond appropriately. Providing face and content validity, and being highly communicative and interactive are obvious advantages to such tasks; while one test taker dominating the interaction, and pairings of test takers with different levels of proficiency are among the disadvantages (O'Sullivan, 2008).

Task Specifications	
Task Input/ Prompt	
Task Content (as informed by the CEFR level descriptions)	B2- C1 CEFR B2: Can use the language fluently, accurately and effectively on a range of general, academic, vocational and leisure topics. CEFR C1: Only a conceptually difficult subject can hinder a natural, smooth flow of language. Can produce clear, well-structured, detailed text on complex subjects. Can give clear, detailed descriptions and presentations on complex subjects, integrating sub-themes. *This task is considered to be beyond B1. At B1 level, the learner "[c]an communicate with some confidence on familiar routine and non-routine matters related to his/her interests and professional field." Although a few of the pictures used in this task may be familiar to the test takers, the five pictures relate to different aspects of a university, which require higher levels of lexical knowledge. Moreover, the task requires complex functions such as expressing and justifying opinions and evaluating the pictures.*
Task Type and Format	The examiner gives the directions for the task and asks the test takers to take an extended turn to complete the whole task.
Task Focus / Target Skills	Negotiation and collaboration skills, goal-oriented cooperation, using functional language to make and respond to suggestions, discuss alternatives, make recommendations and negotiate agreement. The test takers listen and respond to each other, taking turns and making suggestions.
Topic	Negotiating on a university poster
Input	Pictures Instructions directed orally by the examiner
Channel for task delivery	Face to face only, in real time
Interaction pattern	Dialogue: Test taker/test taker
Task instructions	Spoken (real time)
Type of prompt	Pictorial (sequence of pictures) Oral (real time/given by the examiner)
Language level of instructions	Easier than the level of the task
Recorded?	The whole task is recorded for rating
Response	
Response type	Interaction (the two test takers talk to each other)
Timing	5 minutes
Preparation time	0
Interaction required	Interaction with the other test taker

Discourse type required	Discussion/conversation Question and answer
Rhetorical functions	Description (pictures) Suggesting Giving an opinion Asking for opinion Comparing and contrasting Decision making
Register	Informal
Domain	Personal / Educational
Rating Of Task	
Rating method	Analytical method
Rating criteria	Fluency, Range and Accuracy, Lexical Resource, Interaction

Task 4 - Holiday

In this task, the test takers are given four color photographs to describe. The photographs depict people on holiday and the test takers are initially asked to give a simple description of what they see in the photographs. The test takers should be encouraged to give as many details as possible. Then, the examiner asks the test takers to compare and contrast the types of holidays and the people depicted in the photographs. They are asked to talk about why these people chose these kinds of holidays. Finally, the test takers are asked which one of the holiday types shown in the pictures they would prefer, and are expected to extend their answer with reasons and examples. The advantages of tasks that are based on picture prompts include clarity; eliciting extended sample of connected speech; examining test takers' ability to use certain grammatical forms; and valid performance comparisons (O'Sullivan, 2008). The disadvantages of such tasks listed by O'Sullivan (2008) are: limited authenticity; requiring little interaction; unclear and ambiguous pictures; and limited score reliability due to differences in test takers' interpretations of the pictures.

Task Specifications	
Task Input/ Prompt	
Task Content (as informed by the CEFR level descriptions)	A2-B1 CEFR A2: Can give short, basic descriptions of events and activities. Can describe people and places in simple terms. CEFR B1: Can enter unprepared into conversation on familiar topics, express personal opinions and exchange information on topics that are familiar, of personal interest or pertinent to everyday life (e.g. family, hobbies, work, travel and current events). *In this task, the test takers are asked to give simple descriptions of four pictures showing people on different types of holidays. They are expected to describe the places they see (such as the beach, the mountain, or the hotel and the city) and the people (a family, older people, etc.) They are also asked which holiday they would prefer, which is a familiar topic related to travel.*
Task Type and Format	The examiner asks the test takers questions, using standardized questions. The test taker takes relatively longer turns to answer each question.
Task Focus / Target Skills	Describing and commenting on the visual prompt. The test taker gives a description of and responds to questions about the visual prompt. The test taker talks about his/her preferences and support his/her choice with reasons and examples.
Topic	Free time, Entertainment Holidays Human personalities
Input	Pictures Questions directed orally by the examiner
Channel for task delivery	Face to face only, in real time
Interaction pattern	Dialogue: Test taker/examiner Extended replies by the test taker
Task instructions	Spoken (real time)
Type of prompt	Pictorial Oral (real time/given by the examiner)
Language level of instructions	Same as the level of the task
Recorded?	The whole task is recorded for rating
Response	
Response type	Interaction (the examiner asks three questions and the test taker gives extended responses)
Timing	3 minutes
Preparation time	0

Interaction required	Interaction with the examiner Extended responses by the test taker
Discourse type required	Question and answer
Rhetorical functions	Description (pictures) Comparing and contrasting Giving cause-reason
Register	Neutral (neither informal nor formal)
Domain	Personal/ public
Rating Of Task	
Rating method	Analytical method
Rating criteria	Fluency, Range and Accuracy, Lexical Resource, Coherence

Task 5- Extended Talk

In this task, the test takers are given the opportunity to speak for three minutes without interruption. They are given three written prompts and the interlocutor may read the prompts twice, if needed. The test takers are asked to choose one of the topics, prepare for one minute, and talk on the topic they select for three minutes. The test takers have the opportunity to organize their thoughts and ideas, and express themselves coherently in appropriate language. The test takers are asked to pretend as if they were giving a talk in a seminar, addressing an audience, and are given the opportunity to give a talk, with highlighting significant points and giving relevant supporting details. Since the test takers are required to talk for an extended period, they may use a variety of discourse modes and functions. On the other hand, in such tasks some test takers may not find the topics interested, and offering options to test takers to choose from makes comparisons difficult (O'Sullivan, 2008).

Task Specifications	
Task Input/ Prompt	
Task Content (as informed by the CEFR level descriptions)	B2-C2 CEFR B2: Can highlight the personal significance of events and experiences, account for and sustain views clearly by providing relevant explanations and arguments. Can use the language fluently, accurately and effectively on a wide range of academic, vocational or leisure topics, marking clearly the relationship between ideas. Can give clear, systematically developed presentations, with appropriate highlighting of significant points, and relevant supporting detail. CEFR C1: Only a conceptually difficult subject can hinder a natural, smooth flow of language. Can produce clear, well-structured, detailed text on complex subjects. Can give clear, detailed descriptions and presentations on complex subjects, integrating sub-themes, developing particular points and rounding off with an appropriate conclusion. *The topics included in this task are conceptually complex topics. The test takers are required to talk for an extended period of time (three minutes) without interruption from the examiner on the topic they choose. For example, they are asked to talk about the situation of immigrants in their country, as well as explain the problems immigrants face in their country and other countries. They are told that they are talking at a seminar, and are given one-minute preparation time, so that they can give a detailed answer to the question in the three minutes they have.*
Task Type and Format	The examiner gives the directions for the task. The test taker takes one minute to prepare and three minutes to talk.
Task Focus / Target Skills	Describing and making a presentation on a complex, specialized topic; developing particular points that support the test taker's argument, comparing the situations or experiences in two different settings
Topic	Global warming, Socio-economic Status, and Immigrants
Input	Written (the instructions) Instructions directed orally by the examiner
Channel for task delivery	Face to face only, in real time
Interaction pattern	Monologue
Task instructions	Spoken (real time)

Type of prompt	Written and oral instructions (real time/given by the examiner)
Language level of instructions	Easier than the level of the task
Recorded?	The whole task is recorded for rating
Response	
Response type	Monologue (The test taker gives an extended talk)
Timing	3 minutes
Preparation time	1 minute
Interaction required	No interaction
Discourse type required	Speech/ Presentation
Rhetorical functions	Description Referring to past-present Expressing modality Reporting events Speculating Giving cause-reason Giving example Expressing and justifying opinions
Register	Formal
Domain	Educational/Academic
Rating Of Task	
Rating method	Analytical method
Rating criteria	Fluency, Range and Accuracy, Lexical Resource, Coherence

Task 6 - Schedule

In this task, two test takers are presented with a problem-solving (information-gap) task, prompted by visual and written stimuli. Each test taker is provided with a different course schedule, with a list of some elective courses under the schedule. They are asked to talk to each other and decide on two elective courses to take together. The examiner sets up the task, then takes no further part, except in the event of a complete breakdown in the interaction, when the examiner may subtly intervene to redirect the test takers with a further prompt, or when one test taker dominates the interaction, in which case the examiner may encourage the dominated test taker with prompt questions. If one of the test takers dominate the interaction, the examiner is advised to ask facilitative questions to the test taker who had less chance to talk. The test takers are given five minutes to reach agreement and decide on the electives. This task is designed to test negotiation and collaboration skills. The task requires goal-oriented cooperation, and the test takers are expected to follow what is said, repeat back, pass on information, take appropriate turns (initiate, maintain, and end discourse appropriately with effective turn taking), make their opinions understood as regards possible options, giving brief reasons and explanations, and ask for clarification when needed.

Task Specifications	
Task Input/ Prompt	
Task Content (as informed by the CEFR level descriptions)	A2 - B1 CEFR A2: Can interact with reasonable ease in structured situations and short conversations, provided the other person helps if necessary. Can manage simple, routine exchanges without undue effort; can ask and answer questions and exchange ideas and information on familiar topics in predictable everyday situations. Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters to do with work and free time. CEFR B1: Can communicate in simple and routine tasks using simple phrases to ask for and provide things, and to get simple information. Can discuss what to do next, making and responding to suggestions. *This task is considered to be a simple and routine task. The test takers need to ask simple questions to get simple information such as 'Çarşamba günü boş musun?' ('Are you free on Wednesday?'), or 'Hangi saatler boş?' ('Which hours are free'). They need to tell their programs to each other, or make and respond to course suggestions.*
Task Type and Format	The examiner gives the directions for the tasks and asks the test takers to take an extended turn to complete the whole task.
Task Focus / Target Skills	Negotiation and collaboration skills, goal-oriented cooperation; using functional language to make and respond to suggestions, discuss alternatives, make recommendations and negotiate agreement. The test takers listen and respond to each other, taking turns and making suggestions.
Topic	Negotiating on a course schedule
Input	Table (schedule) and program (elective courses) Instructions directed orally by the examiner
Channel for task delivery	Face to face only, in real time
Interaction pattern	Dialogue: Test taker/test taker
Task instructions	Spoken (real time)
Type of prompt	Iconic (table) Textual/written (schedule) Oral (real time/given by the examiner)
Language level of instructions	Same as the level of the task
Recorded?	The whole task is recorded for rating
Response	

Response type	Interaction (the two test takers talk to each other)
Timing	5 minutes
Preparation time	0
Interaction required	Interaction with the other test taker
Discourse type required	Discussion/conversation Question and answer
Rhetorical functions	Suggesting Requesting facts Comparing and contrasting Giving opinions Asking for opinions Decision making
Register	Informal
Domain	Personal
Rating Of Task	
Rating Method	Analytical method
Rating Criteria	Fluency, Range and Accuracy, Lexical Resource, Interaction

APPENDIX C

RATING SCALES

C1 The final form of the scale

FLUENCY

Score	CEFR level	Score Descriptors
11/12	C2	• Expresses him/herself at length with a natural, effortless, unhesitating flow with rare repetition and self-correction • Pauses only reflect search for precision in content
9/10	C1	• Can express him/herself fluently and spontaneously, almost effortlessly with only occasional repetition and self-correction. • Only a conceptually difficult subject can hinder a natural, smooth flow of language.
7/8	B2	• Can communicate spontaneously, often showing remarkable fluency and ease of expression in even longer complex stretches of speech. • May demonstrate language-related hesitation at times, or some repetition and/or self-correction
5/6	B1	• Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production
3/4	A2	• Can make him/herself understood in very short utterances, even though pauses, false starts and reformulations are very evident.
1/2	A1	• Manages very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.

GRAMMATICAL RANGE AND ACCURACY

Score	CEFR level	Score Descriptors
11/12	C2	• Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and eliminate ambiguity. • Maintains consistent grammatical control of complex language.
9/10	C1	• Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly without having to restrict what he/she wants to say. • Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot.
7/8	B2	• Has a sufficient range of language to be able to talk without much conspicuous searching for words, using some complex sentences to do so. • Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding.
5/6	B1	• Uses reasonably accurately a repertoire of frequently used 'routines' and patterns and can correct most of his/her mistakes. • Errors occur, but it is clear what he/she is trying to express.
3/4	A2	• Uses basic sentence patterns with memorized phrases, groups of a few words and formulae • Uses some simple structures correctly, but still systematically makes basic mistakes
1/2	A1	• Has a very basic repertoire of words and simple phrases • Shows only limited control of a few simple grammatical structures and sentence patterns in a learnt repertoire.

LEXICAL RESOURCE

Score	CEFR level	Score Descriptors
11/12	C2	• Has a good command of a very broad lexical; shows awareness of connotative levels of meaning • Consistently correct and appropriate use of vocabulary.
9/10	C1	• Has a good command of a broad lexical repertoire; little obvious searching for expressions and avoidance strategies. • Occasional minor slips, but no significant vocabulary errors.
7/8	B2	• Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution • Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication.
5/6	B1	• Shows good control of elementary vocabulary but major errors still occur when expressing more complex thoughts
3/4	A2	• Can control a narrow repertoire dealing with concrete everyday needs
1/2	A1	• Has a basic vocabulary repertoire of isolated words and phrases.

COHERENCE

Score	CEFR level	Score Descriptors
11/12	C2	• Can create coherent and cohesive text making full and appropriate use of a variety of organizational patterns and a wide range of cohesive devices.
9/10	C1	• Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organizational patterns, connectors and cohesive devices.
7/8	B2	• Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some 'jumpiness' in a long contribution.
5/6	B1	• Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.
3/4	A2	• Can link groups of words with simple connectors like 'and', 'but' and 'because'
1/2	A1	• Can link words or groups of words with very basic linear connectors like 'and' or 'then'.

INTERACTION

Score	CEFR level	Score Descriptors
11/12	C2	• Can interweave his/her contribution into the joint discourse with fully natural turn taking, referencing, allusion making, etc.
9/10	C1	• Can select a suitable phrase from a readily available range of discourse functions to preface his remarks in order to get or to keep the floor and relate his/her contributions skillfully to those of other speakers.
7/8	B2	• Can initiate discourse, take his/her turn when appropriate and end conversation when he/she needs to, though he/she may not always do this elegantly.
5/6	B1	• Can initiate, maintain and close simple conversation. Can repeat back part of what someone has said to confirm mutual understanding.
3/4	A2	• Can answer questions and respond to simple statements. Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own record.
1/2	A1	• Can interact in a simple way but communication is totally dependent on repetition, rephrasing and repair.

0	• No communication possible • No ratable language

C2 Scale development process

McNamara, Hill, and May (2002) point out that the way rating scales and rating criteria are interpreted may "act as *de facto* test constructs" (p. 229). Brown (2005, cited in Ducasse & Brown, 2009) makes a similar point, writing that "in any assessment involving judgment it is the criteria by which the performance is judged which define the construct". The rating scale of a test, therefore, provides a representation of the ideas of test developers regarding the construct being assessed. The rating criteria in the present study are adapted from the CEFR (2001). The level descriptors of the CEFR represent the model of communicative language ability adopted in the CEFR and were developed through both quantitative and qualitative analyses.

The conventional approach to rating scale development has been an a priori approach (Fulcher, 1996). In this approach, the scoring criteria and level descriptions are constructed by experts intuitively. However, Fulcher (1996) favors scale construction based on empirical evidence. Following this line of thought, the initial approach taken in the current study was to collect level descriptors used by various examination boards and those of the CEFR, and adapt them based on an analysis of test taker performances. The level descriptors were drawn from the CEFR, IELTS band descriptors, and Cambridge ESOL examinations. Following the analysis of the audio-recordings and transcripts of student performances, the initial rating scale, which is presented in section C3 below, was developed.

After consultation with an expert from Boğaziçi University, who has extensive experience in language test development and assessment, this initial scale was abandoned. The reason for this was that the performance data obtained in the present study was not large enough to provide a basis for scale development. Since

the development of the current test tasks were informed by the CEFR, only the level descriptors of the CEFR were utilized in the final rating scale. The descriptors in the final form of the rating scale are, therefore, drawn from the CEFR level descriptors only. The current rating scale can and should be refined further through iterations/revisions. In her analysis of interactional competence across proficiency levels, Galaczi (2014) notes that the concepts which the CEFR presents under interactional competence are "undoubtedly at the heart of the IC construct, but they nevertheless need further differentiation between proficiency levels to be usable by relevant stakeholders" (p. 555). This observation may hold true for other criteria included in the current rating scale as well. However, as Weir (2005b) points out, the CEFR level descriptors are "the least arbitrary sequence of scaled proficiency descriptors available to us at the moment" (p. 282).

An important decision that had to be made in the development of the rating scale was whether to develop different scales for each task or use a generic scale for all of the six tasks. Fulcher (1996) argues that most score variance can be explained by test taker ability when rating scales do not refer to specific tasks. Similarly, Fulcher and Reiter (2003) note that when the rating scale is not task specific, ability contributes more to score variance than task conditions. Therefore, the final decision was to use a generic scale that makes no reference to task conditions. Any level descriptors in the CEFR that could refer to task types, task conditions, or tasks were excluded from the scale. For example, descriptors such as the following were either not included in the scale or they were 'cleaned off' of task-specific features to be included in the rating scale:

"… to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics."

“Can produce brief everyday expressions in order to satisfy simple needs of a concrete type: personal details, daily routines, wants and needs, requests for information.”

Scale Descriptors Drawn from Various Sources

Score	Fluency	Accuracy	Range	Coherence	Lexical Resource	Interaction
11/12	Speaks fluently with only rare repetition or self-correction; any hesitation is content-related rather than to find words or grammar Expresses him/herself at length with a natural, effortless, unhesitating flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it Pauses only to reflect on precisely the right words to express his/her thoughts or to find an appropriate example or explanation Varies intonation and stress correctly in order to express finer shades of meaning.	Produces consistently accurate structures apart from 'slips' characteristic of native speaker speech Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions)	Uses a full range of structures naturally and appropriately Exploits a comprehensive and reliable mastery of a wide range of language to formulate thoughts precisely, give emphasis, differentiate and eliminate ambiguity No signs of having to restrict what he/she wants to say	Speaks coherently with fully appropriate cohesive features Develops topics fully and appropriately Creates coherent and cohesive text making full and appropriate use of a variety of organizational patterns and wide range of cohesive devices	Uses vocabulary with full flexibility and precision in all topics Uses idiomatic language naturally and accurately Has a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms Shows awareness of connotative levels of meaning Consistently correct and appropriate use of vocabulary	Interweaves his/her contribution into the joint discourse with fully natural turntaking, referencing, allusion making, etc. Interaction is consistently sustained and developed effectively by initiating and responding appropriately.
.9/10	Speaks fluently with only occasional repetition or self-correction; hesitation is usually content-related and only rarely to search for language Expresses him/herself fluently and spontaneously, almost effortlessly Only a conceptually difficult subject can hinder a natural, smooth flow of language Varies intonation and stress correctly in order to express finer shades of meaning	Produces a majority of error-free sentences with only very occasional inappropriacies or basic/non-systematic errors Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur	Uses a wide range of structures flexibly Selects appropriate formulation from a broad range of language to express him/herself clearly, without having to restrict what he/she wants to say	Produces clear, smoothly flowing, well-structures speech, showing controlled use of organizational patterns, connectors and cohesive devices Speaks at length without noticeable effort or loss of coherence Uses a range of connectives and discourse markers with some flexibility	Uses a wide vocabulary resource readily and flexibly to convey precise meaning Uses less common and idiomatic vocabulary skilfully, with occasional inaccuracies Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions; little obvious searching for expressions and avoidance strategies Occasional minor slips, but no significant vocabulary errors	Interaction is consistently sustained and generally developed effectively by initiating and responding appropriately.

Score	Fluency	Accuracy	Range	Coherence	Lexical Resource	Interaction
7/8	May demonstrate language-related hesitation at times, or some repetition and/or self-correction Communicates spontaneously, often showing remarkable fluency and ease of expression in even longer complex stretches of speech Produces stretches of language with a fairly even tempo; although s/he can be hesitant as s/he searches for patterns and expressions. There are few noticeable long pauses Maintains clear, natural pronunciation and intonation: there are few mistakes but these are rare	Frequently produces error-free sentences, though some grammatical mistakes persist Good grammatical control; occasional 'slips' or non-systematic errors and minor flows in sentence structure may still occur; but they are rare and can often be corrected in retrospect Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes	Uses a range of complex structures with some flexibility Expresses him/herself clearly and without much sign of having to restrict what he/she wants to say Has a sufficient range of language to speak without much conspicuous searching for words, using some complex sentences to do so	Can use a variety of linking words efficiently to mark clearly the relationships between ideas May over-use certain connectives and discourse markers Uses a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some jumpiness in a long contribution	Has a wide enough vocabulary to discuss topics at length and make meaning clear in spite of inappropriacies Uses vocabulary resource flexibly to discuss a variety of topics Has a good range of vocabulary for matters connected to his/her field and most general topics Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution	Initiates discourse, takes his/her turn when appropriate and ends conversation when he/she needs to, though he/she may not always do this appropriately Interaction is sustained in most of the time
.5/6	Usually maintains flow of speech but uses repetition, self-correction and/or slow speech to keep going Produces simple speech fluently but more complex communication causes fluency problems: pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of production Expresses him/herself with relative ease, despite some problems in formulation resulting in pauses A foreign accent is sometimes evident and occasional mispronunciations occur.	Produces basic sentence forms with reasonable accuracy May make frequent mistakes with complex structures, though these rarely cause comprehension problems Communicates with reasonable accuracy in familiar contexts; errors occur, but it is clear what he/she is trying to express	Uses a limited range of complex structures, but these usually contain errors and may cause some comprehension problems Uses a mix of simple and complex structures, but with limited flexibility	Links a series of shorter, discrete simple elements into a connected, linear sequence of points Links basic sentences but with repetitious use of simple connectives and some breakdowns in coherence	Manages to talk about familiar and unfamiliar topics but uses vocabulary with limited flexibility Has a sufficient vocabulary to express him/herself with some circumlocutions on most topics pertinent to his/her everyday life such as family, hobbies and interests, work, travel, and current events Shows good control of elementary vocabulary but major errors still occur when expressing more complex thoughts or handling unfamiliar topics and situations	Often contributes appropriately to the development of interaction; aware of listener; may occasionally rely on interlocutor for continuation of discussion Usually relevant contributions Understands interlocutor; may need clarification at times

Score	Fluency	Accuracy	Range	Coherence	Lexical Resource	Interaction
3/4	Cannot respond without noticeable pauses and may speak slowly, with frequent repetition and self-correction Makes him/herself understood in short contributions, even though pauses, false starts and reformulation are very evident Constructs phrases on familiar topics with sufficient ease to handle short exchanges, despite very noticeable hesitation and false starts Pronunciation is generally clear enough to be understood despite a noticeable foreign accent	Errors are frequent and may lead to misunderstanding Uses some simple structures correctly, but still systematically makes basic mistakes – for example tends to mix up tenses and forget to mark agreement; nevertheless, it is usually clear what he/she is trying to say	Produces basic sentence forms and some correct simple sentences but subordinate structures are rare Has a repertoire of basic language which enables him/her to deal with everyday situations with predictable content, though he/she will generally have to compromise the message and search for words Has a limited repertoire of short memorized phrases covering predictable situations; frequent brekdowns and misunderstandings occur in non-routine situations	Uses the most frequent occurring connectors to link simple sentences in order to tell a story or describe something as a simple list of points Links groups of words with simple connectors like ‘and’, ‘but’ and ‘because’	Is able to talk about familiar topics but can only convey basic meaning on unfamiliar topics and makes frequent errors in word choice Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics Has a sufficient vocabulary for the expression of basic communicative needs Controls a narrow repertoire dealing with concrete everyday needs	Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own accord. May ask and answer simple questions, though he/she will generally requires major prompting and assistance Produces inappropriate and irrelevant responses
.1/2	Pauses lengthily to search for expressions, to articulate less familiar words Little communication possible: cannot participate in true conversational exchange Manages very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication. Pronunciation of a very limited repertoire of learnt words and phrases can be understood with some effort. The interlocutor sometimes needs to ask for repetition	Makes numerous errors except in memorized expressions Shows only limited control of a few simple grammatical structures and sentence patterns in a learnt repertoire	Attempts basic sentence forms but with limited success, or relies on apparently memorized utterances Has a very basic range of simple expressions about personal details and needs of a concrete type	Has limited ability to link simple sentences Gives only simple responses and is frequently unable to convey basic messages Links words or groups of words with very basic linear connectors like ‘and’ or ‘then’	Uses simple vocabulary to convey personal information Has insufficient vocabulary for less familiar topics Has a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations	Can interact in a simple way but communication is totally dependent on repetition, rephrasing and repair Fails to initiate or respond appropriately; causes the breakdown of communication unless misunderstandings are repaired by the interlocutor

C3 The initial form of the scale

FLUENCY

11/12	• Speaks fluently with only rare repetition or self-correction; any hesitation is content-related rather than to find words or grammar • Expresses him/herself at length with a natural, effortless, unhesitating flow
9/10	• Speaks fluently with only occasional repetition or self-correction; hesitation is mostly content-related and only rarely to search for language • Only a conceptually difficult subject can hinder a natural, smooth flow of language
7/8	• May demonstrate language-related hesitation at times, or some repetition and/or self-correction • Produces stretches of language with a fairly even tempo; although s/he can be hesitant as s/he searches for patterns and expressions. There are few noticeable long pauses
5/6	• Usually maintains flow of speech but uses repetition, self-correction and/or slow speech to keep going • Pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of production
3/4	• Cannot respond without noticeable pauses and may speak slowly, with frequent repetition and self-correction • Makes him/herself understood in short contributions, even though pauses, false starts and reformulation are very evident
1/2	• Pauses lengthily to search for expressions, to articulate less familiar words • Manages very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.

GRAMMATICAL RANGE AND ACCURACY

11/12	• Uses a full range of structures naturally and appropriately • No signs of having to restrict what he/she wants to say • Maintains consistent grammatical control of complex language
9/10	• Uses a wide range of structures flexibly • Selects appropriate formulation from a broad range of language to express him/herself clearly, without having to restrict what he/she wants to say • Produces a majority of error-free sentences with only very occasional inappropriacies or basic/non-systematic errors
7/8	• Uses a range of complex structures with some flexibility • Has a sufficient range of language to speak without much sign of having to restrict what he/she wants to say • Frequently produces error-free sentences, though some grammatical mistakes persist
5/6	• Uses a mix of simple and complex structures, but with limited flexibility • May make frequent mistakes with complex structures, though these rarely cause comprehension problems
3/4	• Produces basic sentence forms and some correct simple sentences • Uses some simple structures correctly, but systematically makes basic mistakes
1/2	• Attempts basic sentence forms but with limited success, or relies on apparently memorized utterances • Shows only limited control of a few simple grammatical structures and sentence patterns in a learnt repertoire

LEXICAL RESOURCE

11/12	• Uses vocabulary with full flexibility and precision in all topics • Has a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms • Consistently correct and appropriate use of vocabulary
9/10	• Uses a wide vocabulary resource readily and flexibly to convey precise meaning • Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions; little obvious searching for expressions and avoidance strategies • Occasional minor slips, but no significant vocabulary errors
7/8	• Has a wide enough vocabulary to discuss a variety of topics at length and make meaning clear in spite of occasional inappropriacies • Has a good range of vocabulary for matters connected to his/her field and most general topics • Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution
5/6	• Manages to talk about familiar and unfamiliar topics but uses vocabulary with limited flexibility • Shows good control of elementary vocabulary but major errors and pauses occur when expressing more complex thoughts or handling unfamiliar topics and situations
3/4	• Is able to talk about familiar topics but can only convey basic meaning on unfamiliar topics and makes frequent errors in word choice • Controls a narrow repertoire dealing with concrete everyday needs
1/2	• Uses simple vocabulary to convey personal information • Shows only limited control of a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations

COHERENCE

11/12	• Speaks coherently with fully appropriate cohesive features • Creates coherent and cohesive text making full and appropriate use of a variety of organizational patterns and wide range of cohesive devices
9/10	• Produces clear, smoothly flowing, well-structures speech, showing controlled use of organizational patterns, connectors and cohesive devices • Uses a range of connectives and discourse markers with flexibility
7/8	• Uses a limited number of cohesive devices to link simple and complex utterances into clear, coherent discourse, though s/he may over-use certain connectives and discourse markers
5/6	• Links basic utterances but with repetitious use of simple connectives and some breakdowns in coherence
3/4	• Uses the most frequent occurring connectors to link simple utterances (like ‘and’ [ve], ‘but’ [ama] and ‘because’ [çünkü]
1/2	• Has limited ability to link simple sentences • Links words or groups of words with very basic linear connectors like ‘and’ or ‘then’

INTERACTION

11/12	• Consistently contributes effectively into the joint discourse with fully natural turntaking • Interaction is consistently sustained and developed effectively by initiating, asking and responding appropriately.
9/10	• Maintains and develops the interaction with appropriate turntaking; contributions are appropriate and relevant • Interaction is consistently sustained and generally developed effectively by initiating, asking and responding appropriately.
7/8	• Initiates discourse, takes his/her turn when appropriate, though he/she may not always do this appropriately • Keeps the interaction going with very little prompting and support: contributions are short
5/6	• Contributes to interaction mostly by giving answers to questions posed by the interlocutor • Inappropriate and irrelevant responses occur; but these are rare
3/4	• May answer simple questions, though he/she will generally require major prompting and assistance • Inappropriate and irrelevant responses are very evident; produces very short contributions
1/2	• Communication is totally dependent on repetition, rephrasing and repair • Consistently needs interlocutor support to sustain interaction

0	• No communication possible • No ratable language

APPENDIX D

LEARNER PROFILE FORM

Learner Profile Form

Last Name, First Name:_____

Sex: Female_____ Male:_____

Date of Birth:_____

Place of Birth: City:_____ Country:_____

Mother Tongue: _____

Language of Education:_____

How long have you been learning Turkish? _____

How long have you been living in Turkey? _____

Contact (Mobile phone or email address): _____

Turkish Language Proficiency

How would you rate your linguistic ability in Turkish in the following areas? Please put a tick on the relevant box for each language skill.

	A1	A2	B1	B2	C1	C2
Speaking	I can use simple phrases sentences to describe where I live and people I know. ☐	I can use sentences to describe in simple terms my family, living conditions, my educational background and my present job. ☐	I can describe experiences and events, my dreams, and hopes. I can briefly give reasons and explanations for opinions and plans. I can narrate a story or relate the plot of a book or film and describe my reactions. ☐	I can present clear, detailed descriptions on a wide range of subjects related to my field of interest. I can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options. ☐	I can present clear, detailed descriptions of complex subjects integrating sub-themes. ☐	I can present a clear, smoothly flowing description or argument in a style appropriate to the context and with an effective logical structure. ☐
Interaction	I can interact in a simple way provided the other person is prepared to repeat or rephrase things at a slower rate of speech and help me formulate what I'm trying to say. I can ask and answer simple questions on very familiar topics. ☐	I can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar topics and activities. I can handle very short social exchanges, even though I can't usually understand enough to keep the conversation going myself. ☐	I can deal with most situations. I can enter unprepared into a conversation on topics that are familiar, of personal interest or pertinent to everyday life (e.g., family, hobbies, work, travel, and current events.) ☐	I can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible. I can take an active part in discussions in familiar contexts. ☐	I can express myself fluently and spontaneously without much searching for expressions. I can use language flexibly and effectively for social and professional purposes. I can formulate ideas and opinions with precision. ☐	I can take part effortlessly in any conversation or discussion and have a good familiarity with idiomatic expressions and colloquialisms. I can express myself fluently. If I have a problem I can backtrack and restructure around the difficulty so smoothly that other people are hardly aware of it. ☐

APPENDIX E

TASK EVALUATION FORM

SPEAKING TASK EVALUATION
Task Number:

Surname, Name _____

Class: _____

Were the instructions clear?	YES ☐	NO ☐	If no, which part?

Were the prompts clear?	YES ☐	NO ☐	If no, which part?

Was the time enough to complete the task?	YES ☐	NO ☐	If no, any suggested duration?

Were the pictures helpful in guiding you to speak?	YES ☐	NO ☐

Was the task relevant to what you do (talk about) in real life?	YES ☐	NO ☐

Was the task relevant to what you do in class?	YES ☐	NO ☐

Was the task interesting to you?	YES ☐	NO ☐

About the difficulty level of the task:	Easy ☐	Moderate ☐	Difficult ☐	Very difficult ☐

APPENDIX F

EXEMPLAR PERFORMANCES

Taylor and Galaczi (2011) note that "The aim [in rating scale development] is to describe what learners can do and how well they can do it so that a sample of performance can be matched to a verbal description, usually according to a principle of 'best fit'" (p. 189). The exemplar performances below may not correspond to the verbal descriptors in a one-to-one fashion. In many cases, the level descriptors will not capture the complexity of the language sample produced; and there will be cases where the performance does not feature all the 'component parts' of the descriptors. Each performance, therefore, is matched with the descriptor which it best fits.

Below are the transcription conventions used in the transcriptions of the test taker performances. The transcriptions are more detailed for samples of performances that illustrate Fluency levels.

(0.5)	The number in brackets indicates a time gap in tenths of a second.
[0.5]	The number in squared brackets a time gap including all the non-word sounds which precede it.
hh	An 'h' indicates an out-breath. The more h's the longer the breath.
soun-	A dash indicates the sharp cut-off of the prior word or sound.
sou:::nd	Colons indicate that the speaker has stretched the preceding sound or letter. The more colons the greater the extent of the stretching.
word.	A full stop indicates a stopping fall in tone. It does not necessarily indicate the grammatical end of a sentence.
word,	A comma indicates a 'continuing' intonation.
word?	A question mark indicates rising intonation.
mmm	A 'mmm' indicates interjection (with standardized spelling). It means the speaker makimg 'mmm' sound.

TASK 1: INTERVIEW

GRAMMATICAL RANGE AND ACCURACY

All of the names provided are pseudonyms.

Barbara (12)
Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and eliminate ambiguity.
Maintains consistent grammatical control of complex language.

Yani istanbulu geziyorum. Daha keşfedilecek çok yer var.
Lise döneminde aklımda iki üç meslek vardı. Bunlardan bir tanesi çeviribilimdi, biri psikoloji, biri siyasetti ama kontenjanlar sınırlı olduğu için, siyaset biliminden açılmıştı sadece, hani uluslararası ilişkiler ve siyaseti istiyordum, onu yazdım kazandım. Biraz şansa oldu.
Yani ya siyaset bilimine ya da uluslararası ilişkilere yöneleceğim ama bunların hangisini yapacağıma daha çok emin değilim.
Vizyonda güzel film varsa sinemayı tercih ediyorum çünkü sinemada sinema atmosferi daha iyi film izlemek için. Ama eski zaman- yani eski yapımlı filmleri sinemada bulamadığım için kendi başıma izliyorum.
Burada kalmayacağım. Yani yunanistana da dönmeyebilirim de, burası olmadığını biliyorum zaten mülakat- şey buraya gelirken bi taahhütname imzaladım ve o taahhütnameye göre burada kalamıyorum.
Pozitif yönleri olsa da, eğitim sistemi kötüydü diyebilirim çünkü ne adam akıllı bir türkçe eğitim alabiliyorduk ne adam akıllı bir yunanca eğitim alabiliyorduk. Arada kalıyorduk.
Şimdi kafamda şu vardı, eğer Boğaziçi üniversitesini kazanabilirsem türkiyeye gelirim, boğaziçi olmasa selaniğe giderim. İlk tercihimde de çıkınca buraya geldim.

Natalie (8)
Has a sufficient range of language to be able to talk without much conspicuous searching for words, using some complex sentences to do so.
Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding.

Hiçbir fikrim yok. Yani benim hikayem biraz uzun çünkü yani aslında islam bilimi okudum almanyada… Pek hoşuma gitmedi çünkü islama karşıydı hem de yani pek sevmedim, ondan sonra en kolay şey benim için türkçe, yani Türkçe seçmek. Yani hem kolaydı hem de zaten bundan önce zaten türkçe dersi aldım, yani islam bilimi okurken Türkçe dersi aldım.
Benim birkaç planlarım var ve ben zaten bütün hayatım aynı işi yapmak istemiyorum. Yani öyle düşünüyorum çünkü çok kolay sıkılırım bir işten.
Yani eğlenceli oluyor onlar için. Ondan sonra yani şey, engelli çocuklarla dans etmek tedavi falan vermek istiyorum. Öyle bir planım var.

Ryan (6)
Uses reasonably accurately a repertoire of frequently used ‘routines’ and patterns and can correct most of his/her mistakes.
Errors occur, but it is clear what he/she is trying to express.

Çünkü mesela ben yürüyorken çok istanbulu özellikle çok şeyler görüyorum ve tabi biliyorsam resmi tamlayabilirim.
Hem sosyal medya, face hem iş için, ödev için yapıyorum.
Bazen. Genelde okumuyorum ama bazen çok çalışkan hissediyorsam Türk sayfaları okuyabilirim.
Finans olarak çok büyük bir problem. Çünkü şey her şey emekli olurken devam ettiğini işçiler onlar için emekli insanlar için ödeme- ödemesi lazım.

Eleanor (3)
Uses basic sentence patterns with memorized phrases, groups of a few words and formulae.
Uses some simple structures correctly, but still systematically makes basic mistakes.

Kitap okuyorum. Okuyorum.
Ders çalışıyorum ve yemeği pişi- pişiriyorum.
Arika Jiro yazarı yazdığı kitap seviyorum.
İki kişi kütüphanede çalışıyor. Çok zor.
Dil- dil bilgisi. İngilizce dil bilgisi çalışı- ama Japonya insandan konu- konuşmak istemiyorum. Bu nedenle Japonya Japon insanlar İngilizce konuşa- konuşamaz.

LEXICAL RESOURCE

Barbara (12)
Has a good command of a very broad lexical; shows awareness of connotative levels of meaning.
Consistently correct and appropriate use of vocabulary.

Mübadele sırasında orada kalmışlar ve yunan vatandaşıyız.
Yani konsolos falan olmayı istiyordum böyle büyükelçi falan da, tam daha bölümü bitirmediğim için emin değilim.
Buraya gelirken bi taahhütname imzaladım ve o taahhütnameye göre burada kalamıyorum.
Vizyonda güzel film varsa…
Başka birinin özgürlüğünü kısıtlamadığı sürece…
Düzenli mi diye diyebilirim, derli toplu mu?

Natalie (7)
Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution
Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication.

Türkiyedeki durumu yani engelli durumu pek- şu an yani iyileşiyor. … Engelli çocuklar yani onlara destek veriyorsun.
Ders almasına rağmen bir cümle bile kuramazlar bazıları. Bu çok tuhaf bir şey.
Aslında klasik şeyleri, yani eski zamanlardan, çok severim.
Benim hikayem biraz uzun çünkü yani aslında islam bilimi okudum almanyada.
Aynı zamanda ben zaten türkiyede kalmayı düşünüyorum. İnşallah yani hayırlısı olsa.

Ryan (5)
Shows good control of elementary vocabulary but major errors still occur when expressing more complex thoughts.

Boş zamanımda futbol oynuyorum arkadaşlarımla sohbet ediyorum.
Tabi biliyorsam resmi tamlayabilirim.
Diplomat gibi bir şey olmak istiyorum. Çünkü dünyayı değişebilirim.
Genelde kotu (kötü) bir şey. Amerikada şey birkaç sorunlar var. Altmışlarda çok race sorunlar vardı. Onlar bitmiş ama başka bir şekilde devam ediyor. Haberde çok race rights var.

Eleanor (2)
Has a basic vocabulary repertoire of isolated words and phrases.

Türkçe konu- okuyorum. Ama çok fazla konuşa-
Japonyada İngilizce Japonyada insanlar İngilizce dili bi- sorry grama-

COHERENCE

Barbara (12)
Can create coherent and cohesive text making full and appropriate use of a variety of organizational patterns and a wide range of cohesive devices.

Salsaya gittim, dans kulübüne katılmıştım. Onun dışında takıldığım üç kulüp daha oldu ama şu anda hiçbir şeyle uğraşmıyorum onların dışında.
O konuda tam emin değilim. Yani ya siyaset bilimine ya da uluslararası ilişkilere yöneleceğim ama bunların hangisini yapacağıma daha çok emin değilim.
Pozitif yönleri olsa da, eğitim sistemi kötüydü diyebilirim çünkü ne adam akıllı bir Türkçe eğitim alabiliyorduk ne adam akıllı bi yunanca eğitim alabiliyorduk. Arada kalıyorduk.

The cohesive devices Çisem used: *için, de/da, beri, (onun) dışında, ya… ya da, zaten, göre, çünkü, tabi ki, sürece, bence, öncelikle, veya, gibi, ne… ne, o yüzden, diye, yönelik, özellikle, eğer.*

Natalie (9)
Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organizational patterns, connectors and cohesive devices.

Bu da çok iyi bir soru. Aslında yani şimdi çok anlatacağım ama olsun. Benim birkaç planlarım var ve ben zaten bütün hayatım aynı işi yapmak istemiyorum. Yani öyle düşünüyorum çünkü çok kolay sıkılırım bir işten.
Umut evleri nedir? Orda yani belli bi yaştan sonra mesela on yedi on yaşındayken engelli çocuklar gidiyolar ama çok çok çok zihinsel olmayanlar. Yani onlar kendine karar verebiliyolar. Orda mesela orda engelli çocuklar yani onlara destek veriyosun.

The cohesive devices Jamila used: *ama, yani, için, gibi, yoksa, hem de, bundan önce, ondan sonra, mesela, aynı zamanda, yine de, göre, rağmen.*

Ryan (6)
Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.

Hem sosyal medya face hem iş için ödev için yapıyorum.
Bir tür yok. Sadece aslında mafya filmleri. Mesela goodfellas, casino.
Goodfellas adlı bir film var. Goodfellas'da bir aktör oyuncu var. İsmi Ray Lionel ama filmde ismi Henry Hell. Henry Hell çok hırsız var falan yapıyor. Çok polis onu… onu takip ediyor, ettiği- ettiğinden korkuyor. Sonunda onu buluyor.

The cohesive devices Zachary used: *ama, çünkü, mesela, ve, hem… hem, özellikle, aslında, ya da, sonunda.*

Eleanor (3)
Can link groups of words with simple connectors like 'and', 'but' and 'because'.

Ders çalışıyorum ve yemeği pişiriyorum.
Türkçe okuyorum ama çok fazla konuşa-

The cohesive devices Haruka used: *ve, ama, ya da.*

TASK 2: EVENTS
GRAMMATICAL RANGE AND ACCURACY

Andrew (12)
Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and eliminate ambiguity.
Maintains consistent grammatical control of complex language.

Birinci resimde bir basket maçı var sanırım. Onu onla hakemle teknik direktör tartışıyor gibi görünüyor. Ama basket maçı olduğundan emin değilim. Beyzbol turnu-maçına daha çok benziyor sanırım.
Aşağıdaki resme geçeyim. Aşağıdaki resimde bi doğal afetten sonra insanların çalışmalara yardım ettiğini görüyorum. Sanırım deprem olmuş. Büyük bir bina çökmüş.
Fark gerginliğin farkında olmadıkları için maça odaklanmışlar ve keyif alıyorlar. Buradaki insanlar hayatlarına etki eden bir şey olduğu için hepsi bi el ele verip de iyileşmeye çalışıyorlar. Bi şekilde acıları dindirmeye çalışıyorlar.

Douglas (9)
Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly without having to restrict what he/she wants to say.
Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot.

Diğer resimde dehşet gibi bir şey yani deprem yaşanmış galiba ev komple yıkılmış. Bi de çok insan var, akrabalar bilmem yardım etmek istiyorlar, bir de dehşet altında kalan insanları çıkarmak istiyolar galiba.
Dediğim gibi burda belki yanlış bir karar vermiş. Burda da yani gördüğünüz gibi deprem vardı herhalde, Marmara depremi belki 1999 olmuş.

Brian (7)
Has a sufficient range of language to be able to talk without much conspicuous searching for words, using some complex sentences to do so.
Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding.

İkinci resimde yıkılmış bir ev, belki bomba düşündüğünden sonra ve insanlar akrabalarını arıyorlar.
İkinci resimde çok çok büyük bir dert var insanlarda. Birinci resim ciddi değil yani hiçbir şey, spor yani.
Önce iki takım oynayormuş ve belki bir kavga olmuş ya da bir takımın oyuncusu haksız bir şey olmuş. O yüzden hakim çok kızgınmış kızmış.

Jonathan (4)
Uses basic sentence patterns with memorized phrases, groups of a few words and formulae.
Uses some simple structures correctly, but still systematically makes basic mistakes.

Sebebi bilmiyorum ama basketbol yönetmen çok güzel elbiseler giyebilir.
Bence bu felaket gece gun oldu çünkü hep herkes resimde adam olur ve bence eğer bu sadece bir saattaki iki saatte oldu. Kadınlar resimde olsa çünkü herkes resimde bir adam çünkü onlar belki onlar binası tamir tamir ediyor ediyormuş.

LEXICAL RESOURCE

Andrew (12)
Has a good command of a very broad lexical; shows awareness of connotative levels of meaning
Consistently correct and appropriate use of vocabulary.

İnsanlar göçük altındaki insanları çıkartmaya çalışıyorlar sanırım.
Doğal afetten sonra insanların çalışmalara yardım ettiğini görüyorum.
Bi şekilde acıları dindirmeye çalışıyorlar.
Oyuncuyla teknik direktör ikisi birlik olup şey çullanmışlar yani. Dövecekler.

Douglas (9)
Has a good command of a broad lexical repertoire; little obvious searching for expressions and avoidance strategies.
Occasional minor slips, but no significant vocabulary errors.

İlk resimde üç kişi var. Bir spor salonunda gibi gözüküyor. Spor salonunda bir oyuncu var bir antrenör var bir de hakem var.
Birinci resimde çok adrenalin yani böyle stresli bi şekilde konuşuyolar. İkinci resimde böyle çok yorgun, bir de çok ne diyolar yani şoktalar çok şoktalar.
İkinci resimde nasıl devam etcek? Şimdi yardımlar çoğalcak, mesela itfaiye falan gelecek yardıma.

Brian (7)
Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution.
Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication.

Birincide bir kavga var bu iki sportsmen sporcu sporcu bir şey anlatmak istiyor ve belki kim hakim cevap verene…
Bomba düşündüğünden…
Çok sığınmacılar olacak ama diktatörler yerlerinde kalıyor.

Jonathan (5)
Shows good control of elementary vocabulary but major errors still occur when expressing more complex thoughts.

Genellikle kadınlar insanlar binada felaketin sonra bulunmaz. Bence herkes fotoğrafta mutlu değil tabi ki çünkü felaket felakette oldu.
Hiç kimse mutlu değil bence. Bu adam yanlış şey yanlış şey yapmış ve bu adam onun adı ne?

COHERENCE

Andrew (12)
Can create coherent and cohesive text making full and appropriate use of a variety of organizational patterns and a wide range of cohesive devices.

Birinci resimde bir basket maçı var sanırım… Aşağıdaki resme geçeyim.
Söyledim zaten birincisinde kavga olma ihtimali var. Tartışıyorlar zaten. Daha bi ileri boyuta taşıyabilirler tartışmayı.
Öyle, onun dışında başka bir şey görmüyorum.

The cohesive devices Chasan used: *sanırım, gibi, ama, ve, sonra, onun dışında, için, bence, aynı şekilde, hatta.*

Douglas (10)
Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organizational patterns, connectors and cohesive devices.

İlk resimde üç kişi var, bir spor salonunda gibi gözüküyor… Diğer resimde dehşet gibi bir şey…
Böyle bir şey yok galiba ama bir şekilde ceza vercek hakem.
Dediğim gibi burda belki yanlış bir karar vermiş.

The cohesive devices Mehmet used: *bir de, galiba, belki, herhalde, ama, mesela, gibi.*

Brian (7)
Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some ‘jumpiness’ in a long contribution.

Birincide bir kavga var… İkinci resimde yıkılmış bir ev…
Belki bir kavga olmuş ya da bir takımın oyuncusu haksız bir şey olmuş. O yüzden hakim çok kız kızgınmış.
Şimdi akrabalarını arıyorlar ve yani bundan sonra ne olacak? Bilmiyorum.

The cohesive devices Igor used: *ve, ama, sonra, ya da, o yüzden, sonra.*

Jonathan (5)
Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.

Hakem kurallar anlatıyor. Bu yönetmene anlatıyor ve yönetim yönetmen kızgın çünkü ona göre hakkın yanlış olmuş.

The cohesive devices Richard used: *ve, çünkü, bence, ama.*

TASK 3: POSTER
INTERACTION

Christina (12)
Can interweave his/her contribution into the joint discourse with fully natural turn taking, referencing, allusion making, etc.

Evet ben de ona katılıyorum yani kütüphaneyi insanlar oturuyor.
Evet çok haklısın.
Ben de öyle düşünüyorum.
Hıı evet bak şimdi gördüm bir şey var ödül var orada aşağıda.
Bence gelecek öğrencilere üniversitemizin neler sunduğunu göstermesi gerekiyor.
Hani ikide ve üçte de onu yapıyor, hani üçte hani başarılı olduğumuzu gösteriyor aslında, üç de olabilir başarılı olduğumuz için. Ama yine de sadece teknoloji bölümümüzde başarı değil de diğer bölümler de çok başarılı olduğunu düşünüyorum.

Ryan (7)
Can initiate discourse, take his/her turn when appropriate and end conversation when he/she needs to, though he/she may not always do this elegantly.

Evet bence de ama birinci resimde o kütüphane çok iyi bir kütüphane.
Bence de çok büyük bir fark yok herhangi bir resim ilk ee resimden seçildiğinden fark etmez.
Evet evet kadınlar için bu resim çok, bu resimden o okulu seçmez.

E: *Ee bu bir ekip galiba.*
ZA: *Ekip bilmiyorum.*

Jonathan (4)
Can answer questions and respond to simple statements. Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own record.

KE: *Yani dördüncü koyalım postere ne dersin?*
Rİ: *Ne ne ne dersin?*
KE: *Ne dersin?*
Rİ: *aaa anlamadım.*
KE: *Ne düşünüyorsun?*
Rİ: *Ne ne, dördüncü dördüncü resim hakkında düşünüyorum.*

KE: *Hangi resim? Çünkü bi tane seçmemiz lazım.*
Rİ: *Bana göre galiba birinci birinci fotoğrafta.*

Eleanor (2)
Can interact in a simple way but communication is totally dependent on repetition, rephrasing and repair.

YA: *Deney yapıyorlar belki. Yani üniversitenin laboratuvarları var. Herhalde o anlama geliyor. Üçüncü fotoğraf.*
HT: *A::: ro- robot.*
YA: *Robot, evet evet.*
HT: *Aaa ımmmm.*
YA: *Hangi bölüm olabilir sizce?*
HT: *A:::: ı::: makine makine.*

SAMPLE PERFORMANCES FOR FLUENCY
TASK FA1: INTERVIEW

Barbara (12)
Expresses him/herself at length with a natural, effortless, unhesitating flow with rare repetition and self-correction.
Pauses only reflect search for precision in content.

I: *Neden bu bölümü seçtiniz?*
Çİ: *E::: lise döneminde (0.4) aklımda iki üç meslek vardı. (0.2) Bunlardan bir tanesi çeviribilimdi, biri psikoloji, biri siyasetti. E:: ama kontenjanlar sınırlı olduğu için, (0.2) siyaset biliminden açılmıştı sadece, (0.2) hani uluslararası ilişkiler ve siyaseti istiyordum, onu yazdım kazandım. (0.4) Biraz şansa oldu.*
I: *Peki ileride nasıl bir iş yapmayı düşünüyorsunuz?*
Çİ: *O konuda tam emin değilim. Yani ya siyaset bilimine ya da uluslararası ilişkilere yöneleceğim ama bunların hangisini yapacağıma daha çok (0.2) emin değilim. (0.2) Yani konsolos falan olmayı istiyordum böyle büyükelçi falan da, tam daha bölümü bitirmediğim için emin değilim.*
I: *Burada kalmayı mı yunanistana dönmeyi mi istiyorsunuz?*
Çİ: *Yok burada kalmayacağım. Yani yunanistana da dönmeyebilirim de, (0.2) burası olmadığını biliyorum zaten mülakat (0.2) e:: şey buraya gelirken (0.2) bi taahhütname imzaladım ve o taahhütnameye göre burada kalamıyorum*
I: *Boş zamanlarımda kitap okuyorum dediniz. Ne tür kitaplar okuyorsunuz?*
Çİ: *Psikoloji çok ilgimi çekiyor. mmm politika- (0.2) türk politik tarihine- (0.2) okumaya çalışıyorum çünkü hani tarih bilgim türkiye (0.2) Osmanlı dönemine biraz (0.2) bayağı bi eksiğim var bunları kapatmaya çalışıyorum ve zaten ders materyali çok sayfa olduğu için onlarla da (0.2) onlarla ilgileniyorum.*

Natalie (8)
Communicates spontaneously, often showing remarkable fluency and ease of expression in even longer complex stretches of speech.
May demonstrate language-related hesitation at times, or some repetition and/or self-correction.

JA: *Akşam oraya gidiyorum (0.2) ve: ondan sonr- yani zaten engelli çocuklarda (0.2) yani müzik dans falan ço- çok iyi gidi- gidiyor yani. Yani çok- (0.2) yani eğlence ol- ola- ol- (0.4) eğlenceli oluyor (0.2) onların için. (0.2) Ondan sonra yani*

şey engelli çocuklarla dans etmek (0.2) yani tedavi falan vermek istiyorum. Öyle bi planım var.
I: *Peki kitap okuyor musunuz?*
JA: *Okuyorum. (0.2) Yani türkçe mi soruyorsunuz? Evet ama: (0.6) ya- yani e:: (0.2) okuyorum. Türkçe de okumaya çalışıyorum (0.4) ama (0.2) mesela şu an e:: kocan kadar konuş diye bir (0.2) kitap var. (0.2) Bu çok şey yani eğlencel- yani çok şey. Halit uşaklıgil falan değil yani, bu çok zor (0.2) osmanlıca falan (0.4) ama yani basit şeyleri o- okuyorum.*
I: *Hangi tür okuyorsunuz genelde?*
JA: *Yani as- aslında e:: (0.8) yani mesela ben şu an yani ben yani (1.2) en son o- okuduğum kitap bi de şey halit ziya u- uşaklıgildi ama almancası (0.4) çünkü türkçe çok zor. Ben yani osmanlıca kelimeler falan hiçbir şey anlamıyorum. Ama yani aslında klasik şeyleri yani (0.4) eski zamanlardan (0.2) çok severim.*

Brian (5)
Keeps going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.

I: *Bu kitaplardan bir tanesini bana anlatır mısınız?*
IB: *(1.2) Evet (0.2) e::: mesela (0.6) e:: onu haka::y::-e hikayeleri e:: seyahatlerden e:: anlatı-yor. E::: (0.4) e:: mesela bir hh e::: bir hikaye (0.4) trenle e::: trenle bir (0.2) seyahat e::: s- seyahatten bahsediyor. E::: o::m o::m seyahatin e::: büyüyünce (0.2) o: e: mmm bir e- önceki: bir seyahati e::mm hatırlıyor. (0.8) Ve: e::: (0.6) e:: kar- karşılaş e: e:: karşılaştığı e::: insanları e::: mmm e:: [2.4] insanlardan e::: bahsediyor.*
I: *Evet, peki. Teşekkürler. Peki hangi alanda eğitim alıyorsunuz? Ne okuyorsunuz?*
IB: *E:: Ben e:: mm e: burda araştırma:-yorum. (0.8) E:: Ben (0.2) ders veriyorum e::: tarih bölümünde. (0.4) E::: tarih bölümünde e:::m tarih bölümünde e:: ermeni dili- e::: dilin e::: dersleri e: veriyorum (0.49 ve:: h e:: e- ermeni e:::m e:: ta e:: tarihi e::: tarihin derslerini. (0.2) Ve::: hr- hh doğu (0.4) doğu hristiyanliğin e::: Hristiyanlığın mmm e::: tarihin- e:: tarihin dersle: e::: e:: [1.8] derslerini e:: veriyorum. (0.8) Başka: e:m araştırma e:::m eski edebiyat e::: eski edebiyat hakkında h e::: mm e:: dır. .hhh e:: mm Bizans e:: Ermenistan e e:: e::: suriye e:: ede- edeb- mmm edebiyatlardan.*
I: *Peki doğu Hristiyanlığını bana biraz açıklayabilir misiniz? Bu konuda neler araştırıyorsunuz?*
IB: *E::: >Peki< e:: mm Doğu e:: hristiyan- e:: doğ- doğu hristiyan e:: lıkları e:: diye a:: Bizanstan e:: (0.4) ya da a- ah- Aparatorluktan e:::: do- (0.2) doğuya: (0.2) bulunan e e::: hristiyanl- e::: hristiyan e: ülkeler e::: mm e:: demek. (0.8) E:: Yani kafkaslarda bulunan (0.2) ermenistan, gürcistan e::: sur- e:: suriye e: Kırpt ve: e::: mm e::: e:: [4.2] Habeş-*

Eleanor (2)
Manages very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.

I: *Bazı sorular soracağım sizden cevap vermenizi istiyorum. Evet. Boş zamanlarınızda neler yapıyorsunuz?*
HT: *A::::: (0.4) kitap (0.4) okuyorum. (0.4) Okuyorum.*
I: *Evet başka?*
HT: *A:::::::::: (1.2) ders çalışıyorum a::: ve::::: (1) yemeği (0.4) pişi-(0.2) pişiririm.*
I: *Hangi yemekleri pişiriyorsunuz?*
HT: *A::::: Menemen ve::::: omlet. Immmmmm (0.8) mmm başkaTürk (0.4) Türk yemeği (0.2) pişiririm.*
I: *Peki kitap okumayı seviyorum dediniz. Hangi kitapları okuyorsunuz?*
HT: *A:::::::: (0.4) a:::::: (0.6) a::::::: (0.6) >hangi<?*
…
HT: *A::::::: (0.4) a:::::: (1) Arika Jiro:: yazarı::: (0.4) yazdığı kitap (0.2) seviyorum.*
I: *O kitaptan biraz bahseder misiniz? Anlatır mısınız kitabı bana?*
HT: *A::: .h a:::: ı:::: e:::: a:::::::: [1.2] hangi kitap?*
I: *Sevdiğiniz bir kitap.*
HT: *A::::: e::::: [6.8] iki kişi:: (0.2) kütüphanede:: (0.4) çalışıyo::r a:::::: a::::::: ı:::: [14] Çok zor.*

TASK 4: HOLIDAY
GRAMMATICAL RANGE AND ACCURACY

Sona (11)
Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and eliminate ambiguity.
Maintains consistent grammatical control of complex language.

Dördüncü resim. Burası İsviçre olabilir. Dağlık alan, karla kaplı, ön kısımda yine üç kişi. Keşfe çıkmışlar gibi.
Bi tane seçiliyosa o zaman, bunu derim. Ya ben doğayı çok seviyorum. E coğrafya okuyorum. Yani sonra mesleğim açısından da daha bilinçli olacağımı düşünüyorum. Ya ben ya taşı toprağı seven insanım böyle. Yani deniz olması gerekmiyor benim için, bir yere tatile veya izne gidersem.

Grace (8)
Has a sufficient range of language to be able to talk without much conspicuous searching for words, using some complex sentences to do so.
Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding.

Mesela birincide bence bu resim sonbaharda çekildi. Sonbaharda ağaçlar yapraklar kızarıyor ve çok güzel görünüyor. O yüzden sonbaharda dağa gittiler.

İkincide neden havuzu seçmiş. Yazın çünkü çok sıcak olduğu için ve buradan böyle güzel manzaraya bakarsa çok rahat oluyor o yüzden. Üçüncü de aynı gibi ama bir aile hep beraber eğlenmek için sahil ya da deniz çok güzel bir yer. Dördüncüde bilmiyorum ama belki onlar kayak yapmayı çok sever.

Lee (6)
Uses reasonably accurately a repertoire of frequently used 'routines' and patterns and can correct most of his/her mistakes.
Errors occur, but it is clear what he/she is trying to express.

Birinci fotoğrafta orda üç kişi var. Onlar bir ormanda galiba. Bir dağ var. Sonra bir küçük bir göl var, trekking yapıyorlar.
Bunu seçmeyeceğim çünkü ben İsveçliyim. Sevmiyorum bizim hava. Belki bu. Çünkü newyorka hiçbir zaman gitmedim. Veya bu plac (plaj) evet. Bu placı çok seviyorum.

Alphonso (3)
Uses basic sentence patterns with memorized phrases, groups of a few words and formulae.
Uses some simple structures correctly, but still systematically makes basic mistakes.

Burda ailem yok. O zaman arkadaşlar tatil yapıyorlar. Burda aile var o zaman çocuklar oynamak oynuyorlar ve evet onlar daha eski o zaman ski yapıyorlar.
Benim bu yapacağım çünkü havuzu çok seviyorum ve büyük şehir çok seviyorum.

LEXICAL RESOURCE

Sona (10)
Has a good command of a broad lexical repertoire; little obvious searching for expressions and avoidance strategies.
Occasional minor slips, but no significant vocabulary errors.

Mutlu aile tablosu…
Üç kişi. Herhalde, Türkçesi ne demek, geziye çıktılar gibi.
Keşfe çıkmışlar…
Kilise gözüküyor…
Doğayla iç içe yaşayan insanlar da olabilir.
Varlıklı insanların gidebileceği yerler.

Grace (8)
Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution.
Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication.

Dağa tırmanıyolar, herkes sırt çantası taşıyor.

Şimdi dağda kayak mı, kayak yapıyor.
Ben tek başımda türkiyede oturuyorum ama ailemi çok özlüyorum…

Lee (6)
Shows good control of elementary vocabulary but major errors still occur when expressing more complex thoughts.

Birinci fotografta orda üç kişi var. Onlar bir ormanda galiba. Bir dağ var. Sonra bir küçük bir göl var.
Büyük bir kent ve yüzüyorlar. Evet biraz tehlikli.
Bilmiyorum ben trekking sevmiyorum. Kent kesim, evet.

Alphonso (4)
Can control a narrow repertoire dealing with concrete situations.

Burda aile görüyorum sahilde. Burda üç üç kişi görüyorum dağda evet. Burda üç kişi var dağda.
Bence bu en pahalı çünkü ski her zaman çok pahalı ve sonra bu var çünkü bu havuz otelde ve otele biliyorum ve çok pahalı.

COHERENCE

Sona (11)
Can create coherent and cohesive text making full and appropriate use of a variety of organizational patterns and a wide range of cohesive devices.

Birinci resimde ön kısımda bir göl görüyorum. Arkası dağlık alan. Üç kişi herhalde, türkçesi ne demek, geziye çıktılar gibi.
Veya doğayla iç içe yaşayan insanlar da olabilir ya da ne bileyim coğrafya öğretmenliği okuyan.
Deniz olması gerekmiyor benim için, bir yere tatile veya izne gidersem. Daha çok işte dünyayı böyle keşfetmeyi seviyorum.
The cohesive devices Ayşenur used: *herhalde, veya, ya da, yani, böylece, sanırım, büyük ihtimal.*

Grace (9)
Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organizational patterns, connectors and cohesive devices.

İkincide neden havuzu seçmiş. Yazın çünkü çok sıcak olduğu için ve buradan böyle güzel manzaraya bakarsa çok rahat oluyor o yüzden. Üçüncü de aynı gibi ama bir aile hep beraber eğlenmek için sahil ya da deniz çok güzel bir yer.
The cohesive devices Tomoe used: *ve, mesela, bence, o yüzden, çünkü, ama.*

Lee (7)
Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some ‘jumpiness’ in a long contribution.

Birinci fotografta orda üç kişi var. Onlar bir ormanda galiba.
Üçüncü resimde ne var? Bir aile var, baba anne onların çocukları…
Bunu seçmeyeceğim çünkü ben İsveçliyim. Sevmiyorum bizim hava. Belki bu.
The cohesive devices Alexandre used: *galiba, sonra, ve, çünkü.*

Alphonso (4)
Can link groups of words with simple connectors like 'and', 'but' and 'because'.

Burda üç kişi var dağda ve ski yapıyorlar.
Benim bu yapacağım çünkü havuzu çok seviyorum.
The cohesive devices Patrick used: *ve, ama, o zaman, çünkü.*

TASK 5: EXTENDED TALK
GRAMMATICAL RANGE AND ACCURACY

Sona (11)
Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and eliminate ambiguity.
Maintains consistent grammatical control of complex language.

Almanyada zaten fazla göçmen var. Yani yabancı sayısı çok fazla. En büyük problemleri bence dil konusu. Yani çoğunun almanca bilmemesi ya da genel anlamda o ülke- yani geldikleri ülkenin dillerini bilmeleri. Bu bence eğitimle aşılabilir.
Geldiklerinde ama bazıları işte kendi grupları içinde kalıyor, bazıları da oraya ayak uydurmaya çalışıyor yani asimile olanlar da var bence. Böyle ama kendi hallerinde yani gruplaşmış şekilde kalanlar da var.

Kathy (8)
Has a sufficient range of language to be able to talk without much conspicuous searching for words, using some complex sentences to do so.
Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding.

Şimdi sosyo-ekonomik durum hakkında konuşacağım. Nasıl insanların yaşam tarzlarını nasıl etkiler etkisi olur oluyor. Bence ilk baş ilk olarak eğitimde başlıyor. Ülkeye göre değişebilir mesela biz almanyada çok mesela çok iyi dil öğrenmek öğrenmeyebiliriz çünkü okulda öğreniyoruz ve okul ücretsiz ama başka ülkede belki dil öğrenmek için para çok para olması lazım çünkü özel ders almak gerekiyor.

Kurt (6)
Uses reasonably accurately a repertoire of frequently used 'routines' and patterns and can correct most of his/her mistakes.
Errors occur, but it is clear what he/she is trying to express.

Biraz zamandan sonra norveççe öğreniyorlar ama çünkü norveçte norveçli bilmek çok önemli bir şey. Norveçli biliyorsan işi bula- bulama-
Göçmen mahalleri gibi var sadece. Göçmenler orada yaşıyorlar ve o zaman Norveçli öğrenmek daha zor çünkü Norveç aileleri başka mahallelerde yaşıyorlar. Bu çok çok kötü bence.

Mindy (3)
Uses basic sentence patterns with memorized phrases, groups of a few words and formulae.
Uses some simple structures correctly, but still systematically makes basic mistakes.

Trafik trafik ve bilgisayarı etkileri de insanları çok yorgun yönden yetiyor.
İnsanın yaptığı ettiği işleri bil- şimdi bilgisayarı insan yerine yaptığı için insan rahat da gönderiyor oluy- yaptığı için rahat- insanlar rahat yaşamı…

LEXICAL RESOURCE

Sona (11)
Has a good command of a very broad lexical; shows awareness of connotative levels of meaning.
Consistently correct and appropriate use of vocabulary.

Avrupa sevdası var hepsinin içinde…
Bazıları da oraya ayak uydurmaya çalışıyor yani asimile olanlar da var bence.
Daha samimi bi şekilde yaklaşmak gerekiyor bence. Böyle diyalog olması gerekiyor, iki kültür arasında.

Kathy (8)
Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution.
Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication.

Eğer baba ve anne üniversiteye giderse sonra çocuk mutlaka da üniversiteye gitmeye düşünecek.
Başka etkiler de var. Mesela nasıl ne yiyorlar çocuklar ya da yetişkinler de aynı. Ne yiyorlar, hangi etkinlik seçiyorlar, boş zamanında ne yapıyorlar.
Özel ders almak gerekiyor… Konuşamıyorlar çünkü hiç pratik yapmıyorlardı.

Kurt (6)
Shows good control of elementary vocabulary but major errors still occur when expressing more complex thoughts.

Norveçli biliyorsan işi bula- bulama-. Özellikle belki osloda sorunlar…

Bu güzel bir şey ama güzel bir şey ama tabi ki sorunlar var her yerde gibi norveçte de var… Birkaç Müslüman daha uçta ve bu her yerde görüyoruz ama her yerde gibi her yerde aynı sorunlar var bence.

Mindy (3)
Can control a narrow repertoire dealing with concrete situations.

Bazı insanları rahat rahat yaşamı geçiyor ama teknolojisiz insanlar çok tekniksiz insan insan işçi tekni- insan işçiler doğdu.

COHERENCE

Sona (11)
Can create coherent and cohesive text making full and appropriate use of a variety of organizational patterns and a wide range of cohesive devices.

Şimdi yeni yeni düzenleme yapılıyor. Yeni gelen göçmenlere tabi ki dil testinden geçmeleri gerekiyor, o ülkeye girebilmeleri için. Ya bu bi önlem olabilir, ama bence problemi ortadan kaldırmıyor. Çünkü orda yani senelerce yaşayıp dil bilmeyen insanlar var.

The cohesive devices Ayşenur used: *aslında, ama, yine de, yani, bence, ya da, genel anlamda, genelde, sonra, ne… ne, tabi, tabi ki, için, çünkü.*

Kathy (9)
Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organizational patterns, connectors and cohesive devices.

Şimdi sosyo ekonomik durum hakkında konuşacağım… Sonra başka etkiler de var. Mesela nasıl ne yiyorlar çocuklar ya da yetişkinler de aynı ne ne yiyorlar. Hangi hangi etkinlik seçiyorlar, boş zamanında ne yapıyorlar. Bence daha para daha fazla para olan insanlar belki daha sağlıklı yaşıyorlar çünkü onlar seçebilir.

The cohesive devices Katharina used: *şimdi, bence, ilk olarak, çünkü, eğer, ve, sonra, onun için, ama, mesela, ya da, bu yüzden, o zaman.*

Kurt (7)
Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some ‘jumpiness’ in a long contribution.

Üçüncü konu seçtim. Norveçte çok göçmen göçmen var her yerden ama ilk gelen göçmen türkiyeden ve pakistandan ve tabi ki sorun sorunlar var. Bu göçmenler değişik bir kültür, değişik bir dil, değişik bir din Norveçe getiriyorlar… Tabi ki sorunlar var her yerde gibi norveçte de var. Şu an din sorunları var türkiyede gibi işil ile, kobani ile çünkü çok Müslüman var tabi ki norveçte ve birkaç Müslüman

daha uçta ve bu her yerde görüyoruz ama her yerde gibi her yerde aynı sorunlar var bence.

The cohesive devices Inneborg used: *ama, ve, tabi ki, çünkü, özellikle, o zaman, bence.*

Mindy (3)
Can link groups of words with simple connectors like ‘and’, ‘but’ and ‘because’.

Trafik trafik hakkında sorry ben trafik trafik ve bilgisayarı etkileri de insanları çok yorgun yönden yetiyor… Bazı insanları rahat rahat yaşamı geçiyor ama teknolojisiz insanlar çok tekniksiz insan…

The cohesive devices Mizuki used: *ve, mesela, ama, bu nedenle.*

TASK 6: SCHEDULE
INTERACTION

Amanda (11)
Can interweave his/her contribution into the joint discourse with fully natural turn taking, referencing, allusion making, etc.

Ben de birden önce boşum.
Pazartesi Günümüzde Avrupa Sineması ve işte Shakespeare Okumaları alabiliriz.
Hayır uymuyor.
Diğer dersler de uymuyor pazartesi olan.
Günümüzde Avrupa Sinemasının ne yani ne anlatıyorlar?
Cumhuriyetçilik, Özgürlük ve Demokrasi Okumaları. Hayır ben o dersi almak istemiyorum.
Çünkü aslında politikayı seviyorum.
Ama ders almak için o kadar da yani sevmiyorum.

Grace (8)
Can initiate discourse, take his/her turn when appropriate and end conversation when he/she needs to, though he/she may not always do this elegantly.

Şimdi sen iki tane seçmem seçmemiz gerekiyor.
Şimdi Pazartesi Salı ve Çarşamba var. Sen hangi gün senin için uygun?
Ha, evet benim için de olur olur.
Evet şekspirden avrupa sineması daha ilginç geliyor ben benim için.

I: *Peki siz neden Şekspir okumalarını istemediniz?*
IF: *Çünkü boş*
TK: *Sıkıcı*
IF: *zaman vardı yer vardı.*

Lee (6)
Can initiate, maintain and close simple conversation. Can repeat back part of what someone has said to confirm mutual understanding.

AF: *Bir tane seçtik. Güzel.*
YA: *Peki salı günü ki ders ilginizi çekiyor mu sizin? Cumhuriyetçilik.*
AF: *Evet evet.*
YA: *Özgürlük ve demokrasi okumaları.*
AF: *Evet ilginç bir kurs. Siz ne düşünüyorsunuz?*

YA: *Birlikte iki tane seçmeli ders almamız gerekiyor.*
AF: *Evet, anladım anladım. Evet tamam.*

Mindy (4)
Can answer questions and respond to simple statements. Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own record.

T: *O zaman Şekspir okumaları da uygun değil mi?*
Mİ: a:::
T: *Peki hangisini seçelim? Bana da uygun ikisi de.*
Mİ: *e:: gün günümüzde Avrupa sineması.*

T: *O zaman Cumhuriyetçilik var onu ister misin?*
Mİ: *İstiyorum*
T: *Belki Çarşamba daha ilginç bir ders bulabilir miyiz? Benim Çarşamba neredeyse tamamen boş. Benim ondan on ikiye kadar yoga dersim var. O yüzden yaratıcı yazarlık ve yüzme alabilirim ben.*
Mİ: *Yaratıcı.*
T: *Yaratıcı yazarlık ve yüzme. Diğerleri uymuyor bana. Senin nasıl?*
Mİ: *e:: bir ders var. Çarşamba e günü on ikide on ikiden üçe kadar ticari gazetecilik dersi var.*
T: *Yaratıcı yazarlık ve yüzme sana da uyuyor. Bunlardan hangisini istersin?*
Mİ: *Yüzme yüzme dersi isteri- isterim.*

SAMPLE PERFORMANCES FOR FLUENCY
TASK 5: EXTENDED TALK

Sona (10)
Expresses him/herself fluently and spontaneously, almost effortlessly with only occasional repetition and self-correction.
Only a conceptually difficult subject can hinder a natural, smooth flow of language.

AY: *E:: ben de bir göçmen sa- sayılırım aslında: ı:. Almanya (0.2) doğumluyum ama, (0.4) yine de bilmiyorum, kendimi böyle iki ülke arasında böyle (0.2) git gellerde buluyorum. (0,4) Almanyada (0.2) zaten fazla (0.2) göçmen var. Yanii yabancı sayı- s:: yabancı sayısı çok fazla:. (0.4) E:n büyük problemleri öh öh bence*

e::: dil (0.2) konusu. Yani çoğunun (0.2) almanca bilmemesi (0.4) ya da genel anlamda o ülke- yani geldikleri ülkenin dillerini (0.2) bilmeleri. (0.6) Bu e::: bence eğitimle (0.2) aşılabilir. (1) E:: hh ba:şka (0.6) yani hangi ülkelerden geldikleri belli aslında. (0.4) Fakir ülkelerden geliyorlar. Öh öh, öh öh. Asya afrika gibi. (0.6) Genelde e:::mm böyle bir pardon öh öh e:: (0.4) Avrupa sevdası var hepsinin içinde e içerisinde. Hepsi işte (0.2) para kazanmaya geliyor, daha iyi yaşam şartları için, çalışmak için geliyorlar. (0.4) E:: geldiklerinde ama (0.6) bazıları işte (0.4) kendi grupları ar- e içinde kalıyor, bazıları da oraya a- ayak uydurmaya çalışıyor yani asimile olanlar da var bence. (0.4) Böyle ama kendi hallerinde yani gruplaşmış şekilde (0.2) kalanlar da var. Onların işte en büyük sorunu bence (0.8) kültür farkı oluyor genelde. (0.2) Böyle e:: hh dediğim gibi git-geller oluyor iki kültür arasında. (1.6) Sonra kendini ait hissedemiyor, ne oraya ne buraya. (1.4) Yani işte (1.6) nerden geldiğimiz- geldiğimizi unutmamak lazım ama (0.6) bence insan: kendini nerede iyi hissediyorsa ı oraya ait hissediyor. (1) Yani ben şahsen kendim türküm, almanyada yaşıyorum. Tabi a:: e: türkiye (0.4) vatanım olsa da (0.4) yine de almanyada bir s:: düzenim (0.2) var. Oraya da ait hissediyorum buraya da ait hissediyorum.

Kathy (8)
Communicates spontaneously, often showing remarkable fluency and ease of expression in even longer complex stretches of speech.
May demonstrate language-related hesitation at times, or some repetition and/or self-correction.

KG: *Şimdi sosyo ekonomik durum (0.2) e:: hakkında konuşacağım. (0.2) Nasıl insanla::-rın yaşam tarzları- ta:rzlarını (0.2) etki:l- (0.4) nasıl etkiler (0.2) et-ki::si olur oluyor mmm bence (0.2) i:lk baş:: ilk olarak (0.4) e ö eğitimde başlıyor çünkü:: e:mm eğer baba ve anne (0.2) üniversiteye giders- giderse sonra (0.2) çocuk (0.2) mutlaka da üniversiteye üniversiteg-ye git- gitmeye düşünecek. Belki:: okulda: çok iyi: e- okul onun için biraz zor olabilir ama (0.2) aynı zamanda: daha e:mmm ck (0.4) e::: (1.2) ıhı belki:: (0.6) e::: daha:: sık sık (0.2) üniversiteye (0.2) gider bu çocuklar. Mmm (0.4) sonra: (0.8) başka etkiler de var. mm (0.2) mesela:: e:: nası::l ne- ne yiyo:r yiyorlar çocukla:r ya da e: yetiş-kinler de aynı (0.2) ne: ne yiyorla::r. Hangi e hangi:: et- (0.4) etkinlik seçiyorla:r, boş zamanında ne yapıyorlar. (0.4) Bence e::m (0.4) daha: (0.2) para: daha fazla paha: olan insanla:r (0.4) belki daha sa:ğ-lık-lı sağ- yaşıyorlar (0.2) çünkü:: o- on- onlar seçebilir ve bu yüzden onlar için:: (0.2) onlar daha sa:: şeyler seçebilir. Ehe mmm. Ya da:: mmm (0.6) evet- ve- başka bi:r mesela:: mmm (0.2) ülkeye: (0.2) göre değişebilir mesela biz almanyada çok- e mesela:: e:::m çok iyi (0.2) dil öğrenmek- e öğren-me-ye-biliriz çünkü okulda öğreniyoruz ve okul ücretsiz ama başka ülkede belki (0.2) dil öğrenmek için para: çok para olması lazım çünkü:: mmm (0.4) özel ders almak gerekiyor (0.2) o zaman ülkeye (0.2) göre. Ben burda oku- okula gitmedim, o zaman bilmiyorum ama e: bence (0.2) burda:: e::: (0.4) daha belki derste- çünkü çok e çok türk e: arkadaşım var. Onlar okulda İngilizce öğrendiler ama konuş:- konuşamıyorlar çünkü hiç pratik yapmıyor-lar-dı.*

Lee (6)
Keeps going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.

AF: *Tamam (0.2) tamam e::: o zaman başlıyorum. E::: i:: İsveçte e::: çok (0.2) göç var. (0.2) e::: Çünkü biz h e:::: otus:: otuz yıl-dan beri e::: çok cömert bir e:: göç e:: (0.4) politikamız var. e::: Ve bu yüzden (0.2) çok göç var e:: ülkemde. E::: Olumlu bir- şey- benim için iyi bir şey .hhh e:: İsveçte çok h e::: mmh (0.2)e:: Irak-tan ve bal- e:: bal-kan ülklerden .hhh ve şimdi Suriye-den e:: gelen insanlar var hh e:: .hh Ve:: mmmmmm e: bu da bir e:: çok h e:: politika m e::: göç e::: Kuzey Amerikadan e:: var (0.4) e: Onlar e:: İsveç'in nefusu e::: mmm m ck (1.6) yüz::: on yüz- fen evet (0.6) e::: Tamam İsveçte e:::: hh bizim (0.4) e::: hh en:: büyük problem e:::: (0.2) göç h insanları ile e::: mmm e::: onlar e:: İsveçe e: (0.2) iş e:: bulmak (0.2) çok zor bir şey. Hemen hemen olmaz. h (0.4) Ve e::: bu yüzden onlar e: isveçte mmmb e (0.4) bazen e:: i:: çok e::: yoksul (0.4) ve e:::: bh e onlar e::: kendisi e: dünyada (0.2) yaşıyorlar (0.4) h Ve:: ama e:::: ı:: bd e mesela e:::: mm İran-lı İranlı e:: göç e: insanları. hh onlar e:::: bd iş e:: buldu. Ve::: e:: şimdi çok h e büyük bir problem değil. (0.2) Ve aynı şey e::::: bd e:::: Kuzey Amerikadan göç insanları doğru .hh (0.6) Ama mesela e:::: ı şimdi h e:: Suriyeden e:::: gelen insanlar on- e onlar için (0.2) çok zor. Çok zor. Şimdi e:: çünkü e:::: otuz yıl önce e İsveç (0.2)in ekonomisi (0.4) daha h iyiydi. Hh Şimdi bir e:: ekonomi (0.2) krisis e: var. ve:: herkes içine e:: zor (0.2) bir durum. (0.6) Evet. Hh.*

Mindy (4)
Makes him/herself understood in very short utterances, even though pauses, false starts and reformulations are very evident.

Mİ: *E:::: (1.4) e trafik traf- e::: trafik hakkında:: a:: sorry e:: ben e:: (0.6) mmm (3.2) hh e:: tr- trafik (0.2) trafik e:: ve bilgisayarı (0.6) etkileri (0.2) de e:: (0.4) insanları (0.6) e:: çok yorgun (0.2) yönden-yetiyor ı::: (2.4) ı:: ori- e::: mesela: ı::: (3.2) e: her (0.2) şey bilg- e bilgisayar-ı la (0.8)yan herş (1.2) eskiden (0.6) e:: insan- ı insanın (0.4) yaptığı:- e: ettiği: (0.8) e işleri e:: şimdi: bil- e::: şimdi (0.2) bilgisayarı e:: (0.6) insan yerine (0.2) yaptığı (0.4) için ı::: (3.2) e:: insa::n (2) ı::: (0.8) rahat da (1.4) ı:: gönderiyor- ı::: (0.4) oluy- hhh a:: ı::: yaptığı için (3.8) e:: rahat- e: insanlar rahat e::: yaşamı (2.2) ı:: (1.8) ah bazı insanları (0.2) rahat e:: ins- e:: (0.4) rahat yaşamı (0.4) geçiyor (0.2) ama ı::: (3.8) e:: teko- >tek-no-lo-jis< teknolojisiz (0.2) insanlar ı::: hh (3.4) ı çok iy- tekno- i h tekniksiz insan (1.6) insan i::: işçi ı:: tekni inşa::n e:: işçiler (0.2) doğdu ı::: (2.8) sss ve ı::: (4.2) a:: bu nedenle: e:: (2.2) e:: hep insanları (0.2) mut- (0.2) mutlu değil.*

APPENDIX G

VOCABULARY COUNT

The number of the types and tokens of the words used by ten of the test takers in Form A on Task FA1 and FA2

		Interview (FA1)						Events (FA2)						AllTot
	No of…	Noun	Verb	Adj.	Conj.	Other	Total	Noun	Verb	Adj.	Conj.	Other	Total	
Barbara	Tokens	147	89	54	79	89	458	60	44	36	41	31	212	670
	Types	89	49	31	26	36	231	40	25	20	15	19	119	350
Eleanor	Tokens	42	16	9	6	7	80	6	7	3	0	1	17	97
	Types	24	8	6	3	3	44	5	7	2	0	1	15	59
Andrew	Tokens	103	66	54	45	92	360	60	62	31	22	40	215	575
	Types	67	29	34	20	44	194	31	35	16	11	25	118	312
Brian	Tokens	75	27	21	29	21	173	27	17	23	11	16	94	267
	Types	40	19	11	10	13	93	16	11	12	7	9	55	148
Natalie	Tokens	157	127	60	36	236	616	41	50	25	33	72	221	837
	Types	72	45	24	23	50	214	19	24	8	12	32	95	309
Anna	Tokens	96	90	42	69	142	439	54	34	31	21	58	198	637
	Types	58	35	25	17	46	181	28	18	13	7	24	90	271
Douglas	Tokens	98	75	24	36	57	290	56	45	32	18	41	192	482
	Types	54	36	13	11	25	139	30	25	17	7	17	96	235
Christina	Tokens	115	103	63	55	134	470	41	49	15	17	40	162	632
	Types	65	40	37	16	51	209	28	25	8	10	25	96	305
Jonathan	Tokens	115	44	35	44	51	289	45	19	23	20	25	132	421
	Types	48	18	13	9	19	107	26	11	8	5	12	62	169
Ryan	Tokens	74	54	24	20	47	219	26	32	15	9	19	101	320
	Types	47	25	15	9	24	120	16	15	6	6	10	53	173

The number of the types and tokens of the words used by ten of the test takers in Form B on Task 4 and FB2

		Holiday (FB1)						Extended Talk (FB2)						All Tot
	No of…	Noun	Verb	Adj.	Conj.	Other	Total	Noun	Verb	Adj.	Conj.	Other	Total	
Clint	Tokens	38	35	16	13	32	134	83	52	16	41	69	261	395
	Types	17	16	8	5	16	62	38	19	7	14	22	100	162
Lee	Tokens	35	21	27	10	37	130	48	14	23	20	43	148	278
	Types	27	10	16	4	15	72	21	8	11	5	22	67	139
Sona	Tokens	77	33	25	14	42	191	63	53	19	35	66	236	427
	Types	51	21	17	7	27	123	39	23	11	13	30	116	239
Amanda	Tokens	81	66	30	19	89	285	64	60	33	56	98	311	596
	Types	33	22	15	7	23	100	31	22	20	14	23	110	210
Kurt	Tokens	43	29	27	20	27	146	53	23	27	28	27	158	304
	Types	26	12	14	6	13	71	22	12	16	7	14	71	142
Kathy	Tokens	65	35	49	13	34	196	43	35	16	37	40	171	367
	Types	38	13	17	3	18	89	25	17	12	13	17	84	173
Mindy	Tokens	32	19	14	3	12	80	25	9	9	6	12	61	141
	Types	16	8	12	1	7	44	11	7	5	4	9	36	80
Barbara	Tokens	44	24	17	4	16	105	31	26	12	11	18	98	203
	Types	23	17	9	3	7	59	20	14	8	7	12	61	120
Alphonso	Tokens	16	15	6	11	15	63	21	13	17	14	7	72	135
	Types	10	6	4	3	5	28	18	4	7	5	5	39	67
Grace	Tokens	49	24	23	15	17	128	44	30	23	32	34	163	291
	Types	29	15	11	5	13	73	21	16	14	9	15	75	148

APPENDIX H

SCORES ASSIGNED BY THE FOUR RATERS

FORM 1

Form 1- Rater 1

	T1 F	T1 A	T1 L	T1 C	T1 M	T2 F	T2 A	T2 L	T2 C	T2 M	T3 F	T3 A	T3 L	T3 I	T3 M	Ov. M
Barbara	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12,00
Eleanora	2	3	2	3	2,5	2	3	2	2	2,25	1	3	2	2	2	2,25
Andrew	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12,00
Brian	5	7	6	7	6,25	5	6	7	7	6,25	5	7	7	7	6,5	6,33
Natalie	8	8	7	9	8	9	8	8	9	8,5	9	8	9	10	9	8,50
Anna	10	8	10	9	9,25	10	8	10	10	9,5	10	8	9	11	9,5	9,42
Douglas	11	9	9	10	9,75	11	9	9	10	9,75	10	10	9	11	10	9,83
Christina	12	11	11	12	11,5	11	12	10	12	11,25	11	12	11	12	11,5	11,42
Jonathan	4	4	4	5	4,25	4	4	5	5	4,5	4	5	5	4	4,5	4,42
Ryan	5	6	5	6	5,5	5	5	5	6	5,25	6	6	5	7	6	5,58
Olivia	7	6	6	7	6,5	6	6	6	7	6,25	7	6	6	6	6,25	6,33
Nola	3	3	2	3	2,75	2	3	2	3	2,5	2	3	2	3	2,5	2,58

Note: T1 = Task 1, T2 = Task 2, T3 = Task 3; F = fluency, A = grammatical range and accuracy, L = lexical resource, C = coherence, I = interaction, M = mean, Ov. M = overall mean.

Form 1- Rater 2

	T1 F	T1 A	T1 L	T1 C	T1 M	T2 F	T2 A	T2 L	T2 C	T2 M	T3 F	T3 A	T3 L	T3 I	T3 M	Ov. M
Barbara	12	12	11	12	11,75	12	12	12	12	12	12	12	12	12	12	11,92
Eleanora	3	3	2	2	2,5	2	2	2	2	2	2	3	2	2	2,25	2,25
Andrew	12	12	12	12	12	12	12	12	11	11,75	12	12	12	12	12	11,92
Brian	7	7	6	7	6,75	6	7	7	7	6,75	7	7	7	7	7	6,83
Natalie	9	9	10	10	9,5	9	9	9	10	9,25	10	9	9	9	9,25	9,33
Anna	11	10	11	10	10,5	10	11	10	10	10,25	10	9	10	11	10	10,25
Douglas	11	11	10	11	10,75	11	10	9	10	10	11	10	9	11	10,25	10,33
Christina	12	12	11	12	11,75	12	12	10	12	11,5	12	12	11	12	11,75	11,67
Jonathan	3	3	4	4	3,5	3	2	2	3	2,5	3	4	4	4	3,75	3,25
Ryan	6	6	5	6	5,75	6	5	5	5	5,25	6	6	5	6	5,75	5,58
Olivia	7	6	7	7	6,75	6	6	6	6	6	7	6	6	6	6,25	6,33
Nola	3	3	3	3	3	2	3	2	3	2,5	2	3	3	3	2,75	2,75

Note: T1 = Task 1, T2 = Task 2, T3 = Task 3; F = fluency, A = grammatical range and accuracy, L = lexical resource, C = coherence, I = interaction, M = mean, Ov. M = overall mean.

Form 1- Rater 3

	T1 F	T1 A	T1 L	T1 C	T1 M	T2 F	T2 A	T2 L	T2 C	T2 M	T3 F	T3 A	T3 L	T3 I	T3 M	Ov. M
Barbara	11	12	12	12	11,75	11	12	11	12	11,5	11	12	12	12	11,75	11,67
Eleanora	1	2	2	3	2	2	2	2	3	2,25	1	3	2	3	2,25	2,17
Andrew	10	11	11	11	10,75	11	10	10	11	10,5	10	10	10	11	10,25	10,50
Brian	5	6	6	7	6	5	5	6	6	5,5	5	6	6	7	6	5,83
Natalie	8	7	7	9	7,75	8	7	7	8	7,5	9	8	7	9	8,25	7,83
Anna	9	8	8	9	8,5	9	7	8	9	8,25	8	9	9	10	9	8,58
Douglas	9	8	8	10	8,75	10	9	8	10	9,25	9	9	8	10	9	9,00
Christina	12	10	10	11	10,75	11	10	9	11	10,25	11	10	10	12	10,75	10,58
Jonathan	3	4	4	5	4	4	4	4	4	4	3	4	4	5	4	4,00
Ryan	5	6	5	6	5,5	6	5	4	6	5,25	5	6	5	7	5,75	5,50
Olivia	6	7	6	7	6,5	7	5	6	7	6,25	6	6	6	7	6,25	6,33
Nola	2	3	3	4	3	3	2	3	3	2,75	3	3	3	4	3,25	3,00

Note: T1 = Task 1, T2 = Task 2, T3 = Task 3; F = fluency, A = grammatical range and accuracy, L = lexical resource, C = coherence, I = interaction, M = mean, Ov. M = overall mean.

Form 1- Rater 4

	T1 F	T1 A	T1 L	T1 C	T1 M	T2 F	T2 A	T2 L	T2 C	T2 M	T3 F	T3 A	T3 L	T3 I	T3 M	Ov. M
Barbara	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12,00
Eleanora	2	2	2	3	2,25	2	2	2	2	2	2	2	2	3	2,25	2,17
Andrew	12	12	11	12	11,75	12	11	12	12	11,75	12	12	12	12	12	11,83
Brian	5	6	6	7	6	5	6	6	7	6	6	6	7	7	6,5	6,17
Natalie	9	8	9	9	8,75	9	9	9	10	9,25	10	9	9	10	9,5	9,17
Anna	10	8	10	9	9,25	10	8	10	10	9,5	10	8	9	11	9,5	9,42
Douglas	11	9	9	10	9,75	10	9	9	11	9,75	9	10	9	11	9,75	9,75
Christina	12	11	11	12	11,5	11	10	10	12	10,75	11	12	11	12	11,5	11,25
Jonathan	4	4	4	5	4,25	4	4	4	5	4,25	4	5	5	5	4,75	4,42
Ryan	5	6	5	6	5,5	5	5	6	6	5,5	5	6	5	7	5,75	5,58
Olivia	7	6	7	7	6,75	7	5	6	7	6,25	6	6	6	7	6,25	6,42
Nola	2	3	3	3	2,75	2	2	2	3	2,25	2	3	2	3	2,5	2,50

Note: T1 = Task 1, T2 = Task 2, T3 = Task 3; F = fluency, A = grammatical range and accuracy, L = lexical resource, C = coherence, I = interaction, M = mean, Ov. M = overall mean.

FORM 2

Form 2- Rater 1

	T4 F	T4 A	T4 L	T4 C	T4 M	T5 F	T5 A	T5 L	T5 C	T5 M	T6 F	T6 A	T6 L	T6 C	T6 M	Ov. M
Sona	10	11	10	11	10,5	11	11	11	11	11	11	12	12	12	11,75	11,08
Amanda	10	8	8	10	9	10	9	8	10	9,25	10	10	9	11	10	9,42
Clint	10	9	8	10	9,25	10	8	9	10	9,25	10	10	10	11	10,25	9,58
Kathy	8	8	10	9	8,75	8	8	8	9	8,25	9	9	9	10	9,25	8,75
Grace	6	8	8	9	7,75	7	7	8	8	7,5	7	7	8	8	7,5	7,58
Lee	6	6	6	7	6,25	6	6	5	7	6	5	5	6	6	5,5	5,92
Kurt	5	6	7	7	6,25	5	6	6	7	6	5	6	5	7	5,75	6,00
Rebecca	5	7	7	6	6,25	5	6	6	7	6	7	6	6	7	6,5	6,25
Mindy	4	3	4	4	3,75	3	3	3	3	3	3	4	3	4	3,5	3,42
Alphonso	4	3	4	4	3,75	4	3	4	4	3,75	4	3	3	4	3,5	3,67
Fabio	5	5	6	6	5,5	5	5	5	6	5,25	5	6	6	6	5,75	5,50
Gilbert	6	7	6	7	6,5	6	7	7	7	6,75	7	7	7	7	7	6,75

Note: T1 = Task 1, T2 = Task 2, T3 = Task 3; F = fluency, A = grammatical range and accuracy, L = lexical resource, C = coherence, I = interaction, M = mean, Ov. M = overall mean.

Form 2- Rater 2

	T4 F	T4 A	T4 L	T4 C	T4 M	T5 F	T5 A	T5 L	T5 C	T5 M	T6 F	T6 A	T6 L	T6 C	T6 M	Ov. M
Sona	10	11	11	11	10,75	11	10	10	10	10,25	12	12	11	12	11,75	10,92
Amanda	10	9	9	10	9,5	10	9	8	9	9	10	10	9	11	10	9,50
Clint	10	9	9	10	9,5	10	8	9	9	9	10	10	9	10	9,75	9,42
Kathy	8	8	10	9	8,75	8	8	8	9	8,25	9	9	9	10	9,25	8,75
Grace	6	8	8	9	7,75	7	7	8	8	7,5	7	7	8	8	7,5	7,58
Lee	6	6	5	6	5,75	6	6	5	5	5,5	6	6	6	6	6	5,75
Kurt	5	6	5	6	5,5	5	5	5	6	5,25	5	6	6	6	5,75	5,50
Rebecca	5	7	6	6	6	5	6	6	6	5,75	6	6	7	7	6,5	6,08
Mindy	3	3	3	4	3,25	2	3	3	3	2,75	3	3	3	3	3	3,00
Alphonso	3	3	4	4	3,5	3	3	3	3	3	4	3	4	4	3,75	3,42
Fabio	6	5	6	6	5,75	5	5	5	5	5	6	6	6	6	6	5,58
Gilbert	6	7	6	7	6,5	6	7	6	6	6,25	7	7	6	7	6,75	6,50

Note: T1 = Task 1, T2 = Task 2, T3 = Task 3; F = fluency, A = grammatical range and accuracy, L = lexical resource, C = coherence, I = interaction, M = mean, Ov. M = overall mean.

Form 2- Rater 3

	T4 F	T4 A	T4 L	T4 C	T4 M	T5 F	T5 A	T5 L	T5 C	T5 M	T6 F	T6 A	T6 L	T6 C	T6 M	Ov. M
Sona	10	11	11	12	11	11	10	10	11	10,5	12	12	12	12	12	11,17
Amanda	8	9	10	10	9,25	9	9	8	10	9	10	10	9	11	10	9,42
Clint	9	10	9	11	9,75	9	9	9	10	9,25	11	10	10	11	10,5	9,83
Kathy	8	9	9	10	9	9	8	8	9	8,5	9	9	9	10	9,25	8,92
Grace	7	7	8	8	7,5	7	7	7	8	7,25	7	7	8	9	7,75	7,50
Lee	5	6	6	6	5,75	6	5	5	6	5,5	5	6	6	7	6	5,75
Kurt	5	6	6	7	6	5	5	5	6	5,25	6	7	6	7	6,5	5,92
Rebecca	5	6	6	7	6	5	6	6	7	6	6	7	6	7	6,5	6,17
Mindy	3	3	4	4	3,5	3	3	3	3	3	4	3	3	4	3,5	3,33
Alphonso	4	3	3	4	3,5	3	3	3	4	3,25	4	3	3	4	3,5	3,42
Fabio	5	5	6	6	5,5	5	5	5	5	5	6	5	5	6	5,5	5,33
Gilbert	7	6	6	7	6,5	6	6	6	7	6,25	7	6	7	7	6,75	6,50

Note: T1 = Task 1, T2 = Task 2, T3 = Task 3; F = fluency, A = grammatical range and accuracy, L = lexical resource, C = coherence, I = interaction, M = mean, Ov. M = overall mean.

Form 2- Rater 4

	T4 F	T4 A	T4 L	T4 C	T4 M	T5 F	T5 A	T5 L	T5 C	T5 M	T6 F	T6 A	T6 L	T6 C	T6 M	Ov. M
Sona	11	11	10	11	10,75	11	11	10	11	10,75	12	12	12	12	12	11,17
Amanda	9	8	8	10	8,75	10	9	8	10	9,25	9	10	10	11	10	9,33
Clint	10	9	8	10	9,25	10	8	8	10	9	10	11	10	11	10,5	9,58
Kathy	8	8	8	9	8,25	8	8	8	9	8,25	9	10	9	10	9,5	8,67
Grace	6	8	7	9	7,5	7	7	7	8	7,25	7	8	8	8	7,75	7,50
Lee	6	6	6	7	6,25	6	6	5	7	6	5	5	6	6	5,5	5,92
Kurt	5	6	6	7	6	5	6	5	7	5,75	5	6	6	7	6	5,92
Rebecca	6	7	6	7	6,5	5	6	5	7	5,75	6	6	6	7	6,25	6,17
Mindy	4	3	3	4	3,5	3	3	3	4	3,25	3	4	3	4	3,5	3,42
Alphonso	4	3	4	5	4	4	3	3	4	3,5	4	3	4	4	3,75	3,75
Fabio	5	5	5	6	5,25	5	5	5	6	5,25	5	6	6	6	5,75	5,42
Gilbert	6	7	6	7	6,5	6	7	6	7	6,5	6	7	7	7	6,75	6,58

Note: T1 = Task 1, T2 = Task 2, T3 = Task 3; F = fluency, A = grammatical range and accuracy, L = lexical resource, C = coherence, I = interaction, M = mean, Ov. M = overall mean.

REFERENCES

Açık, F. (2008). *Türkiye'de yabancılara Türkçe öğretilirken karşılaşılan sorunlar ve çözüm önerileri*. Doğu Akdeniz Üniversitesi Eğitim Fakültesi Türkçe Eğitimi Bölümü Uluslararası Türkçe Eğitimi ve Öğretimi Sempozyumu. Retrieved from http://turkoloji.cu.edu.tr/YENI%20TURK%20DILI/fatma_acik_yabancilara_turkce_ogretimi.pdf

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson, & B. North (Eds.), *Language testing in the 1990s* (pp. 71-86). London: Macmillan.

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

Austin, J. L. (1975). *How to do things with words*. Oxford: Oxford University Press.

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition*, *10*, 149-164.

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

Bachman, L. F., & Palmer, A. S. (1981). The construct validation of the FSI oral interview. *Language Learning, 31*, 67–86. doi: 10.1111/j.1467-1770.1981.tb01373.x

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *The Modern Language Journal*, *70*(4), 380-390.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12*(2), 238-257.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86-107.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*. doi: 10.1177/0265532210364404.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, *24*(4), 339-353.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*(1), 1-15.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, *20*(1), 1-25.

Brown, G., & Yule, G. (1983). *Teaching the Spoken Language: An approach based on the analysis of conversational English.* Cambridge: Cambridge University Press.

Brown, G., Anderson, A., Shillcock, R., & Yule, G. (1984). *Teaching Talk: Strategies for production and assessment.* Cambridge: Cambridge University Press.

Brown, J. D. (1999). The relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing*, *16*(2), 217-238.

Brown, J. D. (2005). Statistics corner, questions and answers about language testing statistics: Generalizability and decision studies. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *9*(1), 12–16. Retrieved from http://jalt.org/test/bro_21.htm

Brown, J. D., & Ahn, R. C. (2011). Variables that affect the dependability of L2 pragmatics tests. *Journal of Pragmatics*, *43*(1), 198-217.

Brown, J. D., & Kondo-Brown, K. (2012). Rubric-based scoring of Japanese essays: The effects on generalizability of numbers of raters and categories. In J. D. Brown (Eds.), *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages* (pp. 169-182). National Foreign Language Resource Center: University of Hawai'i at Manoa.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage.* Cambridge: Cambridge University Press.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework*. Princeton, NJ: Educational Testing Service.

Bygate, M. (1987). *Speaking*. Oxford: Oxford University Press.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards, & R. W. Schmidt (Eds.), *Language and communication* (pp. 2-27). New York: Longman.

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

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, *28*(2), 201–219.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, *20*(4), 369-383.

Chalhoub-Deville, M. (2006). Drawing the line: The generalizability and limitations of research in applied linguistics. In M. Chalhoub-Deville, C. A. Chapelle, & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 1-5). Amsterdam: John Benjamin Publishing Company.

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes*, *24*(3), 383-391.

Chambers, F. (1997). What do we mean by fluency? *System*, *25*(4), 535–544.

Clark, J. (1975). Theoretical and technical considerations in oral proficiency testing. In R.L. Jones, & B. Spolsky (Eds.), *Testing language proficiency* (pp. 10-28). Arlington, VA: Center for Applied Linguistics.

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

Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing*, *18*(2), 133-147.

Davies, A. (2010). Test fairness: A response. *Language Testing*, *27*(2), 171-176.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, *26*(3), 367–396.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*(1), 5-34.

Deville, C., & Chalhoub-Deville, M. (2006). Old and new thoughts on test score variability: Implications for reliability and validity. In M. Chalhoub-Deville, C. A. Chapelle, & P. Duff (Eds.), *Inference and generalizability in applied linguistics. Multiple perspectives*. (pp. 9- 25). Amsterdam, the Netherlands: John Benjamin Publishing Company.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing*, *11*(2), 125–144.

Douglas, D., & Selinker, L. (1992). Analyzing oral proficiency test performance in general and specific purpose contexts. *System*, *20*(3), 317-328.

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, *26*(3), 423-443.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289-303.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, 2, 197–221.

Eckes, T. (2009). Many-facet Rasch measurement. *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Retrieved from http://www.coe.int/t/dg4/Linguistic/CEF-refSupp-SectionH.pdf

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt, Germany: Peter Lang.

Educational Testing Service (2004). *IBT/ Next generation TOEFL test independent speaking rubrics (scoring standards)*. Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf

Elder, C., & Wigglesworth, G. (2003). An investigation of the effectiveness and validity of planning time in Part 2 of the IELTS Speaking Test. *IELTS Research Reports, 6*, 1-28.

Elder, C., Iwashita, N. & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer?. *Language Testing, 19*(4), 337-346.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work?. *Language Assessment Quarterly, 2*(3), 175-196.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity and accuracy in L2 oral production. *Applied Linguistics, 30*, 474-509. doi: 10.1093/applin/amp042

Er, O., Biçer, N., & Bozkırlı, K. Ç. (2012). Yabancılara türkçe öğretiminde karşılaşılan sorunların ilgili alan yazını ışığında değerlendirilmesi. *Uluslararası Türkçe Edebiyat Kültür Eğitim Dergisi, 1*(2), 51-69

Field, J. (2011). Cognitive validity. In L. Taylor (Eds.), *Examining Speaking: Research and practice in assessing second language speaking* (pp. 65-111), Studies in Language Testing, 30, Cambridge: Cambridge University Press.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*,299-323.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208-238.

Fulcher, G. (1996b). Testing tasks: Issues in task design and the group oral. *Language Testing*, *13*, 23–51.

Fulcher, G. (2003). *Testing second language speaking*, Harlow: Longman/Pearson Education Ltd.

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

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

Fulcher, G., & Reiter, R. M. (2003). Task difficulty in speaking tests. *Language Testing*, *20*(3), 321-344.

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

Galaczi, E. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests?. *Applied Linguistics*, *35*(5), 553-574.

Galaczi, E., & ffrench, A. (2011). Context validity. In L. Taylor (Eds.), *Examining Speaking: Research and practice in assessing second language speaking* (pp. 112-170), Studies in Language Testing 30, Cambrdige: Cambridge University Press.

Grice, H. P. (1975) Logic and conversation. In P. Cole, & L. Morgan (Eds.), *Syntax and semantics, Vol. 3: Speech acts* (pp. 41-58). New York: Academic Press.

Hasselgreen, A. (2004). *Testing the spoken English of young Norwegians: A study of testing validity and the role of smallwords in contributing to pupils' fluency*. Cambridge: Cambridge University Press.

He, A. W., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young, & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (p. 1-24). John Benjamins Publishing.

Horner, D. (2014). *And what about testing pronunciation? A critical look at the CEFR pronunciation grid and a proposal for improvement*. Retrieved from http://clm-bell.com/wp-content/uploads/2014/02/And-What-about-Testing-Pronunciation.pdf

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it?. *Psychological Methods, 5*, 64–86.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing*, *17*(3), 123-139.

Hubbard, C., Gilbert, S., & Pidcock, J. (2006). Assessment processes in speaking tests: A pilot verbal protocol study. *Research Notes, 24*, 14- 19

Hughes, A. (1989). *Tests for language teachers*. Cambridge: Cambridge University Press.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

International English Language Testing System (2009). *SPEAKING: Band descriptors (public version)*, retrieved from https://www.ielts.org/pdf/SpeakingBanddescriptors.pdf

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135-159.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct?. *Applied Linguistics*, *29*(1), 24–49.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to test design. *Language Learning, 51*(3), 401-436.

Jamieson, J. M., Eignor, D., Grabe, W., & Kunnan, A. J. (2008). Frameworks for a new TOEFL. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language™* (pp. 55-95). Routledge.

Jeong, H., Hashizume, H., Sugiura, M., Sassa, Y., Yokoyama, S., Shiozaki, S., & Kawashima, R. (2011). Testing second language oral proficiency in direct and semidirect settings: A social-cognitive neuroscience perspective. *Language Learning*, *61*(3), 675-699.

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485-505.

Kane, M. (2010). Validity and fairness. *Language Testing*, *27*(2), 177-182.

Kang, O. (2008). Ratings of l2 oral performance in English: Relative impact of rater characteristics and acoustic measures of accentedness. *Spaan Fellow. Working Papers 6*, 181-205.

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

Khalifa, H., & Weir, C. J. (2009). *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing 26, Cambridge: Cambridge University Press.

Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, *26*(2), 187-217.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, *70*(4), 366-372.

Lado, R. (1961). *Language testing.* London: Longman.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*, 33–51.

Lee, Y.W. (2005). *Dependability of scores for a new ESL speaking test: Evaluating prototype tasks*. TOEFL® Monograph MS-28. Princeton, NJ: ETS.

Lee, Y. W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, *23*(2), 131-166.

Lee, Y. W., & Kantor, R. (2005). Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes. *ETS Research Report Series*, *2005*(1), i-76.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language learning*, *40*(3), 387-417.

Levelt, W. J. M. (1989). *Speaking.* Cambridge, Massachusetts, London: MIT Press.

Lewis, M., & Starks D. (1997). Revisiting examination questions in tertiary academic writing. *English for Specific Purposes, 16*, 197-210.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54-71.

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing*, *22*(4), 415-437.

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

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158-180.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, *26*, 397–422.

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, *8*(2), 127-145.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

McNamara, T. F. (1997). "Interaction" in second language performance assessment: Whose performance?. *Applied Linguistics*, *18*, 444–446.

McNamara, T.F. & Adams, R.J. (1991). Exploring rater behavior with Rasch Techniques. Paper presented at the Annual Language Testing Research Colloquium (Princeton, NJ, March 21–23). Eric document ED345 498.

McNamara, T., Hill, K., & May, L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics*, *22*, 221-242.

McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell Publishing.

Mehnert, U. (1998). The effects of different length of time for planning on second language performance. *Studies in Second Language Acquisition*, *20*, 83-108.

Mehrang, F., & Rahimpour, M. (2010). The impact of task structure and planning conditions on oral performance of EFL learners. *Procedia-Social and Behavioral Sciences*, *2*(2), 3678-3686.

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

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning*, *49*, 285-310.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189-227.

Nakatsuhara, F. (2008). Inter-interviewer variation in oral interview tests. *ELT Journal*, *62*(3), 266-275.

Norton, J. (2005). The paired format in the Cambridge Speaking Tests. *ELT Journal*, *59*(4), 287-297.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, *19*(3), 277-295.

O'Sullivan, B. (2008). Notes on assessing speaking. Retrieved from http://lrc.cornell.edu/events/past/2008-2009/papers08/osull1.pdf

O'Sullivan B., & Green, A. (2011). Test taker characteristics. In L. Taylor (Eds.), *Examining Speaking: Research and practice in assessing second language speaking* (pp. 36-64), Studies in Language Testing 30, Cambridge: Cambridge University Press.

O'Sullivan, B., Weir, C., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing 19*(1), 33-56.

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition, 21*, 109-148.

Politt, A., & Murray, N. L. (1996). What do raters really pay attention to?. In M. Milanovic, & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium* (pp. 74–91). Cambridge: Cambridge University Press.

Porter, D. (1991). Affective factors in language testing. In J.C. Alderson, & B. North (Eds.), *Language testing in the 1990s*. Modern English Publications in association with The British Council. London: Macmillan, 32-40.

Purpura, J. E. (2008). Assessing communicative language ability: Models and their components. In *Encyclopedia of language and education* (pp. 2198-2213). Springer US.

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

Ross, S., & Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition*, *14*(2), 159-176.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing* 24(3), 355–90.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922.

Shaw, S. D. & Weir, C. J. (2007). *Examining Writing: Research and Practice in Assessing Second Language Writing.* Studies in Language Testing 26, Cambridge: UCLES/Cambridge University Press.

Shi, L. (2001). Native-and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*(3), 303-325.

Skehan, P. (1998). *A Cognitive Approach to Language Learning.* Oxford: Oxford University Press.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510-532. doi:10.1093/applin/amp047.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, *9*, 239-261.

Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing*, *18*, 275–302.

Tarone, E. (1998). Research on interlanguage variation: Implications for language testing. In L.F. Bachman, & A.D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 71-89). Cambridge: Cambridge University Press.

Taylor, L. (2001). The paired speaking test format: Recent studies. *Cambridge ESOL Research Notes, 6*, 15-17.

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

Taylor, L., & Galaczi, E. (2011). Scoring validity. In L. Taylor (Eds.), *Examining Speaking: Research and practice in assessing second language speaking* (pp. 171-233), Studies in Language Testing 30, Cambridge: Cambridge University Press.

Taylor, L., & Wigglesworth, G. (2009). Are two heads better than one? Pair work in L2 assessment contexts. *Language Testing*, *26*, 325–339.

Teng, H. (2007). *A study of task types for L2 speaking assessment*. (ERIC Document Reproduction Service No. ED496075). Retrieved from http://eric.ed.gov/?id=ED496075

University of Cambridge ESOL Examinations (2015). *Information for candidates*. Retrived from http://www.cambridgeenglish.org/images/173976-cambridge-english-advanced-examiners-comments.pdf

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263-287.

Weir, C. J. (1993). *Understanding and developing language tests.* NewYork: Prentice Hall.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach,* Basingstoke: Palgrave Macmillan.

Weir, C. J. (2005b). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, *22*(3), 281-300.

Weir, C. J., & Wu, J. R. (2006). Establishing test form and individual task comparability: A case study of a semi-direct speaking test. *Language Testing*, *23*(2), 167-197.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*(3), 305-319.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*, 85-106.

Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly, 7*(1), 1-24.

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231-252.

Xi, X., & Mollaun, P. (2006). Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST). *ETS Research Report Series*, i-71. doi: 10.1002/j.2333-8504.2006.tb02013.x

Xi, X., & Mollaun, P. (2009). How do raters from India perform in scoring the TOEFL iBT™ speaking section and what kind of training helps?. *ETS Research Report Series*, *2009*(2), i-37.

Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning*, *61*(4), 1222-1255.

Young, S., & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition, 14*(4), 403-424.

Yuan, F., & Ellis, R. (2003). The effects of pretask planning and on-line planning on fluency, complexity, and accuracy in L2 monologic oral production. *Applied Linguistics*, *24*, 1–27.

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs?. *Language Testing, 28*(1), 1-20. doi:10.1177/0265532209360671.