

THE DEVELOPMENT OF A LISTENING TEST FOR LEARNERS OF
TURKISH AS A FOREIGN LANGUAGE

EMEL TOZLU

BOĞAZİÇİ UNIVERSITY

2017

THE DEVELOPMENT OF A LISTENING TEST FOR LEARNERS OF
TURKISH AS A FOREIGN LANGUAGE

Thesis submitted to the
Institute for Graduate Studies in Social Sciences
in partial fulfillment of the requirements of the degree of

Master of Arts
in
English Language Education

by
Emel Tozlu

Boğaziçi University

2017

DECLARATION OF ORIGINALITY

I, Emel Tozlu, 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 09/02/2017

ABSTRACT

The Development of a Listening Test for Learners of Turkish as a Foreign Language

The aim of this study is to report the development and validation process of a listening test for learners of Turkish as a foreign language (TFL). The test tasks developed for this study ranged from A1 to B2 levels according to the Common European Framework of Reference for languages (CEFR, Council of Europe, 2001) and the test was administered in two pilot sessions to learners of TFL at Boğaziçi University. Weir's (2005) socio-cognitive framework for validating language tests and Field's (2013) model of listening comprehension were the two major frameworks adopted for test validation and development in this study. For investigation of validity, the test tasks and the test takers' responses to the test tasks were analyzed in terms of three essential components of Weir's framework, i.e. theory-based validity, context validity and scoring validity. Theory-based validity is examined according to the cognitive requirements specified in Field's listening model and the CEFR descriptors for listening. Contextual features of the tasks were scrutinized based on the contextual parameters outlined by Weir (2005). Investigation of scoring validity of the test takers' responses was conducted through classical item analysis procedures, i.e. central tendency measures, reliability and item analysis. In addition to these analyses, the task evaluation questionnaires given the participants for each task provided valuable quantitative data for cognitive, context and scoring validity. In the light of discussions provided throughout the study based on qualitative and quantitative data, suggestions for the future versions of the test and further research were also mentioned.

ÖZET

Yabancı Dil Olarak Türkçe Dinleme Sınavı Geliştirilmesi

Bu araştırmanın amacı, Türkçe'yi yabancı dil olarak öğrenen kişilerin dinleme becerilerini ölçen bir testin geliştirme ve geçerlilik çalışması sürecini rapor etmektir. Bu çalışma için geliştirilen sınav görevleri, Diller için Avrupa Ortak Öneriler Çerçevesi (Common European Framework of Reference for Languages (CEFR), Avrupa Konseyi, 2001) referans alınarak A1 seviyesinden B2 seviyesine kadar değişiklik göstermiştir ve hazırlanan sınav iki pilot deneme halinde Boğaziçi Üniversitesi'nde Türkçe öğrenen yabancı öğrencilere uygulanmıştır. Bu çalışmada esas olarak alınan iki kuramsal çerçeve, Weir'ın (2005) dil sınavlarının geçerliliğini ölçen sosyo-bilişsel çerçevesi ve Field'ın (2013) dinleme algılama modelidir. Geçerlilik araştırması için, sınav görevleri ve sınav katılımcılarının bunlara cevapları, Weir'ın sosyo-bilişsel geçerlilik çerçevesinin üç önemli ögesi- kuram-bazlı geçerlilik, bağlamsal geçerlilik ve puanlama geçerliliği- açısından analiz edilmiştir. Kuram-bazlı geçerlilik, Field'ın dinleme modelinde ve CEFR'daki dinleme becerisi tanımlayıcılarında belirtilen bilişsel istemlere göre incelenmiştir. Test görevlerinin bağlamsal özellikleri, Weir tarafından belirtilen bağlamsal unsurlara göre ayrıntılı bir şekilde incelenmiştir. Katılımcıların verdikleri cevapların puanlama geçerlilik çalışması da, klasik madde analizi yöntemleri- genel eğilim ölçümleri, güvenilirlik ve madde analizi- yoluyla yapılmıştır. Tüm bu analizlere ek olarak, katılımcılara her görev için verilen görev değerlendirme anketleri de kuram-bazlı, bağlamsal ve puanlama geçerlilikleri açısından değerli nicel veri sağlamıştır. Çalışma boyunca nicel ve nitel veriler baz alınarak yapılan tartışmaların ışığında, testin gelecek versiyonları ve gelecek araştırmalar için önerilerden de bahsedilmiştir.

ACKNOWLEDGEMENTS

I, firstly, would like to thank my thesis advisor, Assist. Prof. Aylin Ünaldı, for her invaluable support throughout my thesis. No word can express my gratitude for her guidance, patience, friendship and encouragement throughout this difficult journey. I always felt very lucky to work with her and to get to know her both as an academician and a person. I also would like to thank Assoc. Prof. Gülcan Erçetin, who encouraged me to study in this area and made valuable contributions to my thesis. I also feel grateful to Assist. Prof. Zeynep Banu Koçoğlu for her valuable comments on my study. My gratitude also goes to Prof. Ayşe Gürel for inspiring me to be a part of an important project like this one. I also appreciate Z. Ceyda Arslan Kechriotis' help and support for data collection and her important feedback on my tasks.

My dearest friends, Esra Keskin Korumaz, Büşra Tombak İlhan and Gözde Büklüm also deserve a big 'thank you' for their continuous emotional support and friendship. Thank you for all the hugs, laughs and love! I should also thank my friend, Koray Tunç for his valuable technological support and friendship.

I owe the biggest thanks to my lovely family, who were always encouraging, supportive and understanding, and to my beloved husband, Uğur Tozlu, who always helped and supported me in my good and bad times. Without their precious love and support, I could not have gone through this period as easily. Their unconditional love and confidence in me always gave me the power to finish what I started. Thank you for being in my life and making me "me"!

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION.....	1
1.1 Introduction to the study.....	1
1.2 Aims of the study.....	2
1.3 Research questions	3
1.4 Organization of the thesis.....	4
CHAPTER 2: LITERATURE REVIEW.....	6
2.1 Introduction.....	6
2.2 Validity.....	7
2.3 Weir's socio-cognitive framework.....	10
2.4 Test development.....	46
2.5 Conclusion.....	53
CHAPTER 3: METHODOLOGY.....	55
3.1 Introduction.....	55
3.2 Participants.....	55
3.3 Instruments.....	57
3.4 Procedure.....	59
3.5 Data analysis.....	61
3.6 Conclusion.....	69
CHAPTER 4: RESULTS AND CONCLUSION.....	70
4.1 Introduction.....	70
4.2 Investigation of theory-based validity.....	70
4.3 Investigation of context validity.....	98
4.4 Investigation of scoring validity.....	131

4.5 Conclusion.....	153
CHAPTER 5: CONCLUSION.....	154
5.1 Summary of the findings.....	154
5.2 Limitations of the study.....	156
5.3 Suggestions for further research.....	158
APPENDIX A: CEFR DESCRIPTORS FOR OVERALL LISTENING	
COMPREHENSION.....	160
APPENDIX B: TASKS FOR THE FIRST PILOTING.....	161
APPENDIX C: TEST TASKS FOR THE SECOND PILOTING.....	176
APPENDIX D: TEST SPECIFICATIONS FOR THE SECOND	
ADMINISTRATION.....	192
APPENDIX E: TASK EVALUATION QUESTIONNAIRES.....	200
APPENDIX F: CONSENT FORM.....	204
APPENDIX G: CEFR DESCRIPTORS FOR VOCABULARY KNOWLEDGE...	206
APPENDIX H: FUNCTIONAL DIMENSIONS OF THE LISTENING TEXTS...	208
APPENDIX I: ORDER OF THE MEAN SCORES IN THE FIRST AND SECOND	
PILOT ADMINISTRATIONS.....	213
REFERENCES.....	215

LIST OF TABLES

Table 1. Lower-level Test Takers' Perceptions of Cognitive Processes in A1 Level	
Task.....	74
Table 2. Lower-level Test Takers' Perceptions of Cognitive Processes in A2 Level	
Task.....	81
Table 3. Higher-level Test Takers' Perceptions of Cognitive Processes in A2 Level	
Task.....	83
Table 4. Lower-level Test Takers' Perceptions of Cognitive Processes in B1 Level	
Task.....	87
Table 5. Higher-level Test Takers' Perceptions of Cognitive Processes in B1 Level	
Task.....	88
Table 6. Higher-level Test Takers' Perceptions of Cognitive Processes in B2 Level	
Task.....	94
Table 7. Test Takers' Perceptions of the Clarity of the Instructions.....	102
Table 8. Sample Responses and Results for Spelling Mistakes.....	106
Table 9. Ratio of Timings to Items across Tasks.....	108
Table 10. Test Takers' Perceptions of the Times Recordings are Heard.....	111
Table 11. Text Purpose and Discourse Modes Across Tasks.....	113
Table 12. Text Lengths across Different Tasks.....	114
Table 13. Topics in the Texts and Their Compatibility with the CEFR	
Descriptors.....	116
Table 14. Test Takers' Perceptions of the Recorded Texts in terms of	
Relevance.....	117
Table 15. Test Takers' Perceptions of Text Difficulty.....	124

Table 16. Speech Rates across Different Tasks.....	125
Table 17. Mean Scores for the Speed of the Recordings.....	126
Table 18. Test Takers' Perceptions of the Comprehensibility of the Texts.....	127
Table 19. Test Takers' Perceptions of the Audibility of the Texts.....	128
Table 20. Descriptive Statistics of the Total Test Scores from the First Pilot Administration.....	132
Table 21. Mean Scores for the Tasks in the First Pilot Administration.....	133
Table 22. Item Analysis Statistics in the First Pilot Administration.....	135
Table 23. Descriptive Statistics of the Test Scores in the Second Pilot Administration for Lower-level Test Takers.....	138
Table 24. Mean Scores for the Tasks in the Second Pilot Administration for Lower- level Test Takers.....	139
Table 25. Test Takers' Perceptions of Task Difficulty in the Second Administration.....	140
Table 26. Item Analysis Statistics for the Lower-Level Test Takers in the Second Pilot Administration.....	142
Table 27. Descriptive Statistics of the Test Scores in the Second Pilot Administration for Higher-level Test Takers.....	148
Table 28. Mean Scores for the Tasks in the Second Pilot Administration for Higher- level Test Takers.....	149
Table 29. Item Analysis Statistics for the Higher-Level Test Takers in the Second Pilot Administration.....	150
Table 30. Item Analysis Statistics for the Hypothetical Test Scores from the Second Administration.....	151

LIST OF APPENDIX TABLES

Table H1. Functional Dimensions of the Listening Texts.....	208
Table I2. Order of the Mean Scores of the Items in the First Pilot Administration.....	213
Table I3. Order of the Mean Scores of the Items in the Second Pilot Administration for Lower-level Test Takers.....	214
Table I4. Order of the Mean Scores of the Items in the Second Pilot Administration for Higher-level Test Takers	214

LIST OF FIGURES

Figure 1. A socio-cognitive framework for validating listening tests.....	14
Figure 2. Aspects of context validity for listening.....	29
Figure 3. Types of listening as determined by listener's goals.....	31
Figure 4. A system of text purposes.....	38
Figure 5. A scheme of test development.....	48
Figure 6. Distribution of the test scores in the first pilot administration.....	133
Figure 7. Distribution of the test scores in the second administration for the lower- level test takers.....	138
Figure 8. Distribution of the test scores in the second administration for the higher- level test takers.....	148

LIST OF APPENDIX FIGURES

Figure G1. CEFR descriptors for vocabulary knowledge	206
--	-----

CHAPTER 1

INTRODUCTION

1.1 Introduction to the study

This study is based on the development of a listening test for learners of Turkish as a foreign language (TFL). Validity and reliability are two essential facets of test development. In general, test validation entails that test developers need to collect evidence for the validity and reliability of the claims and inferences that they make about test performances of test takers. Collecting evidence for validity and reliability of inferences should be dependent on a test validation framework so that the process of gathering evidence will be systematic and based on theory. In this study, Weir's (2005) socio-cognitive framework for validating language tests has been adopted. This framework focuses on the investigation of test taker characteristics, theory-based (cognitive, or construct) validity, contextual validity, scoring validity (reliability), consequential validity and criterion-related validity. The current study only explores three components of Weir's (2005) framework; namely, theory-based validity, context validity and scoring validity.

Furthermore, during test development, we also need to define the construct under investigation according to a theoretical framework. Field's (2013) framework for listening comprehension and the Common European Framework of Reference for Languages: learning, teaching and assessment (CEFR, the Council of Europe, 2001) were two essential theoretical frameworks adopted for the operationalization of the listening construct in this study. These frameworks are integrated into the test validation process via the discussion of theory-based validity in Weir's socio-cognitive framework.

1.2 Aims of the study

Turkish as a foreign language is becoming more and more popular and the number of people learning Turkish to use it in their daily lives or in academic settings is increasing day by day. This situation has caused a need for more standardized tests to assess Turkish language proficiency as a foreign language. A few Turkish proficiency tests are offered to foreign language learners of Turkish. For instance, Turkish Proficiency Test (TYS) delivered by Yunus Emre Institute Exam Center consists of sub-tests that measure reading, writing, listening and speaking. The tests are given in paper-and-pencil test format. Test takers are given a certificate indicating their scores and the predicted CEFR proficiency level. The European Language Certificates (TELC) also administers Turkish listening tests to foreigners in both general and educational domains and the tests are aligned with the CEFR levels. Moreover, Distance Turkish Test (UTS) devised by TÖMER-Ankara is an online test, which assesses six different components of languages such as listening, grammar, reading, speaking interaction, speaking production and writing. These tests are widely used in the assessment of Turkish language proficiency. The present test is offered as another TFL test but one which specifically focuses on academically-related language use. The present study also exemplifies a rigorous attempt of building theoretical and statistical evidence collected through a test validation framework and a listening framework. The validation study conducted for this listening test provides a strong theoretical basis for the development of a listening test. Therefore, this study aims to provide a well-designed, theoretically strong and systematic listening test for learners of TFL. Another main aim of this study is to offer a listening test that can be used as a proficiency or placement test for the Turkish language courses at Boğaziçi University. Every year many Erasmus students

and other overseas learners who come to Boğaziçi University take Turkish courses. However, these students are not placed into proficiency levels based on a standardized assessment, but rather according to students' self-evaluations, teachers' evaluations or through a non-standardized test prepared by the teachers at the institution. With the help of the current test the listening proficiency levels of Turkish learners at Boğaziçi University will be determined in a more standardized and reliable way. The tasks and the topics in the present test are designed according to the needs of university students. Therefore, this test is specifically considered suitable for assessing the Turkish proficiency levels of university students. This listening test is a part of a larger project which also consists of reading (Kurt, 2015), speaking (Gülle, 2015) and writing (Küçük, 2017) components. With each of these components combined, a comprehensive proficiency test of Turkish will have been developed.

1.3 Research questions

In the light of the information given above, three main research questions have been formulated in this study. The research questions that we aim to answer in this study are as follows:

1. What are the cognitive requirements of the listening test tasks?
 - a. Is the listening construct operationalized in the test tasks in a way that targets a sufficient range of cognitive processes indicated by the listening frameworks across different proficiency levels as predicted by the CEFR?
 - b. Do the test takers' perceptions of the listening sub-skills that they employed to answer the items support that the test tasks can tap into the necessary cognitive processes?

2. What are the contextual characteristics of the listening test tasks?
 - a. What are the demands imposed upon the test takers by task setting, administration setting, linguistic features of the listening test tasks and the speakers?
 - b. What are the participants' perceptions of the tasks in terms of the suitability of their contextual features for the different proficiency levels?
3. How well do the test tasks and the items function in terms of scoring validity?
 - a. Do the values for central tendency measures of the tasks and item analyses based on the test takers' performances support that the test is functioning well?
 - b. Does the test measure the listening ability of learners of TFL reliably?

The first question investigates the cognitive requirements of the test tasks and their appropriacy and adequacy according to Field's (2013) listening comprehension framework, the CEFR descriptors for the listening skill and the test takers' perceptions of the listening sub-skills. The second research question seeks answers for the contextual parameters of the test tasks following Weir (2005) and Elliott and Wilson's (2013) discussions on context validity and through the test takers' perceptions of the test tasks based on the answers that they gave in the task evaluation questionnaires. The last question is related to the reliability aspect of test development and statistical analyses are conducted to answer this question.

1.4 Organization of the thesis

This thesis is comprised of the following chapters: Chapter 2 offers a detailed literature review on the crucial aspects of test development; i.e. a general discussion

on validity, the test validation framework by Weir (2005), the listening model proposed by Field (2013) and test development steps to follow based on Weir's (2005) test validation framework. Chapter 3 gives a comprehensive outline of the participants, instruments, data collection procedures and data analysis methods employed to examine the three research questions in this study. Chapter 4 demonstrates a detailed investigation of the research questions with the analyses of the test tasks in terms of theory-based validity, context validity and scoring validity. In Chapter 5, the findings are summarized, and limitations of the current study and suggestions for future research are stated.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Test validation is an on-going process of providing evidence that the scores obtained from a language test reflect the language construct we are attempting to measure and possibly nothing else (Bachman & Palmer, 1996). Test performance should be a clear reflection of the actual use of the construct in the target language use (TLU) domain so that we can ensure that the generalizations that we make to the TLU domain based on test performance are valid and reliable. In each step of test development, we aim to provide evidence related to the two essential aspects of test development, i.e. validity and reliability. Accordingly, since this is a test development study, the issues of validity and reliability (or scoring validity) are discussed thoroughly in this chapter. In the following pages, a test validation framework that forms the theoretical basis of the test validation process in this study is also presented. In addition, the target language skill or construct should also be examined in order to give a clear picture of what the test aims to assess. Therefore, in this chapter, theories and frameworks suggested to explain the listening construct are discussed, and a specific listening framework that theoretically underlies the listening test under investigation is scrutinized in order to make explicit how the construct is defined in this study. In addition to this framework, the CEFR is also discussed briefly since the listening descriptors in the CEFR are also used to define the construct in this study. Finally, after presenting the validation framework and the listening framework, certain steps are suggested for test development and the links between validation and test development process are underlined.

2.2 Validity

Two views of validity, classical and modern, are mainly adopted by researchers. The old notion of validity was emphasized in classical test theory. For example, Kelley (1927, in Weir, 2005, p.12) states ‘the problem of validity is that of whether a test really measures what it purports to measure.’ Similarly, Lado (1961, in Weir, 2005, p.12) maintains ‘Does a test measure what it is supposed to measure? If it does, it is valid.’ These definitions of validity according to classical test theory demonstrate that validity is closely related with whether what we are trying to measure is what the test is supposed to measure.

However, later on, a more modern interpretation of validity was adopted, especially after Cronbach’s (1971) emphasis on the validity of the interpretations of the test scores, not the test scores themselves (in Secolsky, Buchanan & Drane, 2015). Messick (1989, in Messick, 1993, p.1) defines validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment”. Similarly, Weir (2005) defines validity from a more modern perspective and writes:

Validity is perhaps better defined as the extent to which a test can be shown to produce data, i.e., test scores, which are an accurate representation of a candidate’s level of language knowledge or skills. In this revision, validity resides in the scores on a particular administration of a test rather than in the test per se. (p.12)

Similar to Weir (2005), Douglas (2010) explains what validity is about in the following way:

. . . collecting evidence to demonstrate that the interpretations and decisions we make on the basis of test performance are justified. In order to do this, we need to focus on the ability(ies) the test is intended to measure and the decisions we wish to make on the basis of it, and then collect evidence in support of our claim that the test does what we intend it to do. (p.26)

From these definitions it can be understood that there is a shift of focus from the validity of the test itself to the validity of the scores, and the interpretations and inferences made based on those scores. As Weir (2005) explains, it is not a test that is valid or invalid, but the inferences based on the test scores produced as a result of a particular administration of a test on a particular group of test takers. He adds that if different versions and administrations of a test yield consistent scores throughout years, it can be concluded that it is a valid test over time and across different administrations and samples. However, Bachman (1990) argues test validation is not concerned with investigating the validity of the content of the test or of the test scores; instead, it examines the validity of how the information collected via the test is interpreted and used. This shows clearly that validity is fundamentally related to the interpretation of the test scores.

Validation may begin before the operational use of the test and in ideal terms it should continue as long as the test is used and new issues related to validity arise. Justifying uses of the tests we give is crucial since the decisions we make are likely to have an impact on individuals, programs, institutions, or organizations. Therefore, considering the impact of an assessment in many ways, we need to make preparations to be accountable in terms of the decisions or inferences we make about the test takers based on their performances in the test. We can achieve this by collecting evidence that supports our decisions and test use (McNamara, 2000).

The accuracy of the measurement, the extent to which the test measures what it aims to measure, and evidence for the justification of the decisions based on the test scores are key elements of test validity (Douglas, 2010, p.26). McNamara (2000) states that problems related to crucial aspects of tests can pose a threat to validity. These essential areas include test method, test content, and test construct (p.50).

Douglas and McNamara like other researchers agree on the key parts of validity. Therefore, investigation of validity related to these aspects of tests is of utmost importance.

Bachman (1990) states that traditionally validity is divided into categories such as content, construct and criterion validity. However, he continues that a unitary view of validity has been embraced by measurement specialists. Messick (1993) argues that the traditional types of validity are limiting in some ways and they are not alternatives to each other; instead, they complement one another. Therefore, a unitary concept of validity has been widely accepted today. Similarly, Bachman (1990) also maintains that although it is still necessary to collect evidence for these different types of validity, they do not suffice to show the validity of a specific interpretation or use of test scores; rather, all types of information need to be gathered to demonstrate validity. In addition, according to Weir (2005), none of these different types of validity can be considered superior to one another. All dimensions of validity are important and none can be ignored, as problems in only one can affect test scores and interpretation. He also argues that different types of validity are directly linked to each other. Contextual factors have an impact on the cognitive processes employed during listening and the marking criteria used during scoring (scoring validity). For example, the task type that is chosen for a test influences the cognitive processes that are elicited and also has a big effect on the way the responses are assessed. In other words, one aspect of the test has implications for the other aspects, too. Therefore, for a thorough validation study, all components of validity should be taken into account and supported with evidence. They cannot be considered as separate entities (Bachman, 1990; Messick, 1993; Weir, 2005).

Furthermore, Weir (2005) questions the traditional a posteriori evidence collection methods and argues that a priori evidence should also be collected. Weir states that a priori evidence has generally been neglected and a posteriori evidence collection has been deemed sufficient for test validation; however, for a posteriori evidence we still need to determine the target skill and sub-skills, and define the construct; therefore, doing this before the administration would be more helpful for test givers. Weir (2005) reflects his ideas regarding the importance of a priori evidence as follows:

There is a need for validation at the a priori stage of test development. The more fully we are able to describe the construct we are attempting to measure at the a priori stage the more meaningful might be the statistical procedures contributing to construct validation that can subsequently be applied to the results of the test. Statistical data do not in themselves generate conceptual labels. We can never escape from the need to define what is being measured, just as we are obliged to investigate how adequate a test is in operation. (p.18)

Consequently, we can argue that different types of validity require different types of evidence collection; therefore, it is important to gather information both before and after the test if we would like to carry out a comprehensive validation study.

2.3 Weir's socio-cognitive validity framework

Numerous frameworks for validating language tests have been developed by researchers. One of them is a quite comprehensive framework proposed by Weir (2005). He offers a socio-cognitive framework for test validation and the steps that need to be followed for validity and reliability concerns. The essential components to be investigated in Weir's (2005) socio-cognitive framework are as follows:

- Test taker characteristics
- Theory-based validity
- Context validity
- Scoring validity
- Criterion-based validity
- Consequential validity

Firstly, test taker characteristics are emphasized in Weir's framework and it reflects the social aspect of it. Weir states that physical, psychological and experiential differences of the test takers should be taken into account while tests are being developed so that no individual test taker will be put at a disadvantage due to such differences. Secondly, theory-based validity means that a test needs to correlate greatly with behaviors that can be expected based on theory, but also it should not have a significant correlation with other variables with which correlation is not expected. In addition, context validity, typically named as content validity, is concerned with the contextual properties of the language tests and their appropriacy for measuring language ability. Moreover, scoring validity is a superordinate term used by Weir (2005) to refer to all aspects of reliability. He argues that scoring validity is related to the extent to which the test scores are free from measurement error and therefore the extent to which test developers can rely on them while making inferences about test takers according to their performances. Furthermore, consequential validity is integrated into Weir's framework and it is relevant to the sources of bias in a test, its impact on teaching and learning, and its effects on society. Finally, criterion-related validity is connected with the relationship between a test and an external criterion that is considered to measure the same language ability, as well as the extent to which the test predicts future performances of test takers. As previously mentioned, the current study only focuses on the theory-based, context and scoring validity components of Weir's framework and therefore, these three will be explained below in more details. The other aspects of the framework- individual characteristics, consequential validity and criterion-related validity- will not be investigated in the current study; therefore, they will not be discussed any further as part of this study.

Before a detailed explanation of theory-based, context and scoring validity, I would like to present an overall picture of what Weir's socio-cognitive framework entails and the reason why this study is based on this framework. It should also be noted here that Weir (2005) described the components of his framework individually for each language skill and the important parameters identified for validating a listening test were taken into consideration in this validation study. Therefore, the readers are advised to keep in mind that the following discussion concerns the components of the framework for listening tests rather than all language skills.

Geranpayeh and Taylor (2008) maintain that Weir's socio-cognitive framework is composed of three important dimensions- internal cognitive processing, external contextual factors, and individual characteristics. This shows that this socio-cognitive framework takes into consideration the mental processes included in the comprehension processes, the context in which the listening event and test administration happen, and the listeners as a social being and the sources of knowledge they bring with them. Taylor (2013) explains that internal cognitive processing includes processing the acoustic-phonetic input and the grammatical structure, making inferences and self-monitoring as well as applying other sources of information such as language and content knowledge. There are also certain external contextual factors that shape the listening comprehension process such as the setting, the demands of the listening task on the listener and other variables related to the acoustic input. In addition, the individual characteristics of the listener also play a significant role in the comprehension process. They bring physical, psychological and experiential characteristics into the process, which inevitably shapes the results of the listening event. All these combined we can see a test validation model that takes into account both the cognitive and social side of the listening process.

Taylor (2013) also maintains that Weir's framework provides us with "a theoretically grounded and empirically oriented cognitive processing framework" to analyze and understand the different dimensions of listening tests in terms of different validity types (p.25). All components of the socio-cognitive framework for validating listening tests can be seen in Figure 1 (Weir, 2005, p.45). This figure shows how each aspect of validity is linked to one another and how the validation process works for listening tests. Weir also argues that some essential questions, which are listed below, must be addressed by test developers and users while validating a listening test via the socio-cognitive framework (2005, pp.48-9):

1. How are the physical/physiological, psychological and experiential characteristics of candidates addressed by the test? [Test taker characteristics]
2. Are the contextual characteristics of the test task and its administration situationally fair to the candidates? [Context validity]
3. Are the cognitive processes required to complete the tasks interactionally authentic? [Theory-based validity]
4. How far can we depend on the scores on the test? [Scoring validity]
5. What impact does the test have on its various stakeholders? [Consequential validity]
6. What external evidence is there that the test is doing a good job? [Criterion-related validity]

Seeking answers for these questions guides test developers and helps them gather validity evidence in a systematic and multidimensional manner; therefore, this framework is considered to be appropriate for the aims of this study. As previously mentioned, the cognitive processes elicited from the test takers by the test tasks, the contextual characteristics of the test tasks and test administration, and the extent to which the test scores can be depended upon are the main concerns to be investigated in the current study. Thus, they will be scrutinized further in the following sections.

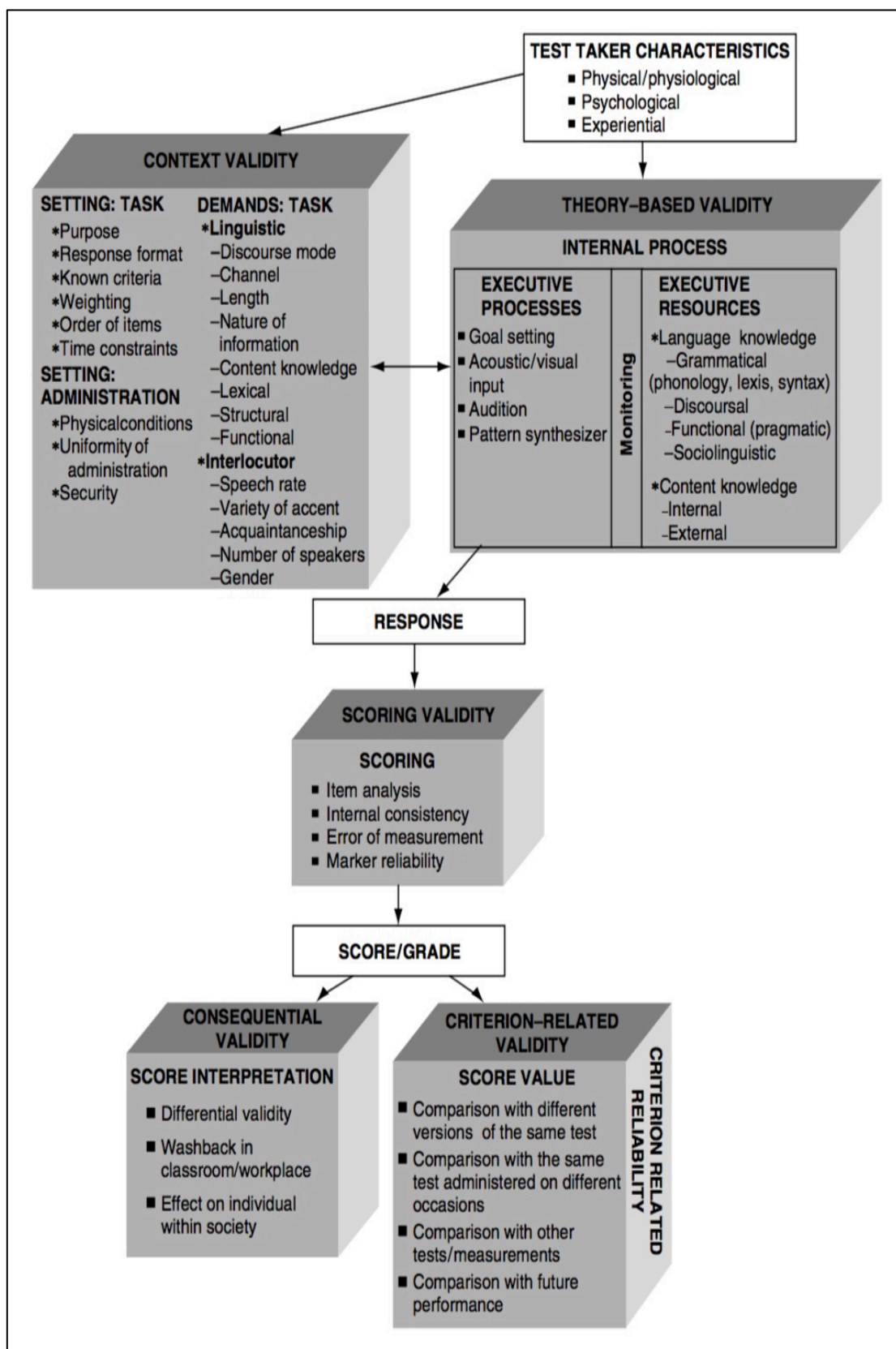


Figure 1. A socio-cognitive framework for validating listening tests (Weir, 2005, p.45)

2.3.1 Theory-based validity

Theory-based validity, construct validity or as later named in further work (Khalifa & Weir, 2009) cognitive validity, is one of the criteria of Weir's (2005) socio-cognitive framework for validating language tests. For theory-based validity, Field maintains "A strand of construct validity, it addresses the extent to which a test requires a candidate to engage in cognitive processes that resemble or parallel to those that would be employed in non-test circumstances" (2013, p.78). In this statement, the focus is on the cognitive processes elicited from test takers by test tasks and real-world situations in which listening happens. Field (2013) also discusses that the important point is not whether the test is close to an actual speaking or listening event, but whether the test tasks require test takers to employ the mental processes that a normal language user would do in a target real-world situation. Eliciting adequate representatives of these mental processes is the main concern for cognitive validity. According to Field (2013), representativeness issue deals with three main points: similarity of processing, comprehensiveness and calibration. This means that the test tasks should elicit a broad range of processes (comprehensiveness) that are similar to those in real-life (similarity of processing) and the cognitive demands of the tasks should be level-appropriate in terms of the performance demands of each proficiency level (calibration).

Weir (2005) suggests that evidence for theory-based validity can be collected both before the administration of the test, e.g. via reports from expert opinions, and after the administration through statistical analyses of the scores obtained from the test and criterion-related studies. However, the most important component of a priori evidence collection is the evidence gathered about the target construct. We need to demonstrate that the test measures the 'listening construct' and nothing else if

possible. Bachman and Palmer (2010) state that in order to be able to avoid any interference by the individual attributes, the language abilities we are aiming to measure, or the construct, should be defined precisely. Construct is defined for assessment purposes as “the specific definition of an ability that provides the basis for a given assessment or assessment task and for interpreting scores derived from this task.” (Bachman & Palmer, 2010, p.43). In order to be able to justify the inferences we make based on the test scores, we must define the construct first and then we need to show evidence that the test, the tasks and the test scores are all related to our construct as well as the test purpose (Douglas, 2010). Similarly, Bachman and Palmer (1996) point out that construct definition is necessary and useful in terms of its role as a basis for utilizing the test scores for their intended purpose(s), as a guide in the test development process and as evidence for the cognitive validity of the interpretations made from the test scores (p.117). Thus, it can be stated that our construct definition affects the whole test.

Another issue related to constructs is what threatens them. Theory-based validity can be threatened by two factors: ‘construct under-representation’ and ‘construct irrelevant variance’ (Messick, 1989, in Messick, 1993). Construct representation means that the test needs to cover the important aspects of the construct sufficiently and should not be too narrow (Messick, 1993); otherwise, the validity claims are threatened by construct under-representation. For instance, assessing only phoneme or word level comprehension in a listening test at a high proficiency level would cause this threat, as it fails to include essential dimensions of the listening construct such as lexical or sentence level comprehension or discourse construction. Another threat to construct validity, construct irrelevant variance entails that test givers need to make sure that the test measures the construct(s) that it

intends to and other irrelevant factors to the construct such as test method, bias, etc. are not at play (Messick, 1993). McNamara (2000) points out that asking a student questions that require having knowledge of a particular topic instead of the construct leads to construct irrelevant variance since the questions assess something other than the language construct. This can work for the advantage or disadvantage of the student depending on the existence or the lack of the knowledge on the part of the student. Therefore, it is essential that test designers pay paramount attention to the possibility of these threats and take precautions against them.

In brief, it can be stated that defining the construct and determining its components are crucial to establish theory-based validity and also context validity. In the following pages, firstly, a number of frameworks defining the listening construct are investigated, as the adequacy and accuracy of the definitions of listening construct forms the theoretical basis for our theory-based and context validity claims. Secondly, a five-stage listening framework from a cognitive perspective proposed by Field (2013) is explained in detail since the listening construct in this study is defined according to this framework. Finally, the CEFR is also discussed shortly to demonstrate how the listening descriptors in the CEFR are useful in defining the listening construct.

2.3.2 Frameworks for the listening skill

Field (2013) argues that listening has not been given much importance in second language research as opposed to the recognition that it has got in first language speech science and psycholinguistics. He also maintains that most research on listening has focused on speech perception and not much interest has been shown to what happens after the listeners receive the signal and how they handle the message.

Listening comprehension is an unobservable event that occurs in the listener's mind; therefore, it is essential to understand the cognitive processes employed during listening comprehension. In order to shed light on the nature of the listening process, several frameworks emphasized differing levels of separation among listening sub-skills.

One of the early views regarding listening processing is "the two-stage view" (Buck, 2001, p.51). Many researchers supported this view and made a distinction between a first step where the input is processed in terms of linguistic information and a second step where the linguistic knowledge is processed for wider meanings and they also argue that these two stages are not necessarily in a linear order, but in an interactive process. (Clark & Clark, 1977, in Buck, 2001).

Another cognitive model of listening comprehension was proposed by Valette (1977, in Buck, 2001). According to Valette's taxonomy, there are five levels that show the "complex cognitive skills" that can be employed during listening comprehension. These levels are mechanical skills, knowledge of the language, transfer, communication and criticism.

Furthermore, Rost (2013) mentions a model of listening comprehension by Demyankov (1983), which is composed of six stages of language understanding (Rost, 2013, p.6):

1. acquisition of the linguistic framework of the language in question;
2. construction and verification of hypothetical interpretations of what is heard;
3. discernment of the speaker's intentions;
4. assimilation of the spoken message;
5. coordination of the speaker's and listener's motivation for participation in the conversation;
6. discernment of the tone of the message.

Rost (2013) discusses that although this model seems to include many important aspects of listening ability, it does not reflect the nature of real-time listening event

due to its stages. Rost also emphasizes that the listener does not have to apply these stages in a linear order or follow all of them in order to have a sensible understanding of the input.

Moreover, Cutler and Clifton (1999) proposes a listening model which consists of four levels; which are decoding, segmenting, recognizing and integrating. This model makes a distinction between sound and word recognition and divides the process into two as “decode and segment” (Field, 2013). However, it does not focus much on meaning construction, but rather on speech perception and analysis.

Nation and Newton (2009) argue that although listening was mostly seen as a passive process in the past, nowadays, most researchers see it as a complex, interactive and interpretive process in which listeners process the input, and form meanings and interpretations based on the input and their background knowledge. A model of language comprehension reflecting such a view was proposed by Anderson (2000). Anderson (2000) makes a distinction between three stages of processing during language comprehension; namely, perception (decoding), parsing and utilization stages. In terms of the listening skill, the first stage, perception, includes perceptual processes that help listeners to decode the spoken message. The second stage of listening, parsing, refers to the process during which listeners convert the words in the spoken message into “a mental representation of the combined meaning of the words” (Anderson, 2000, p.313). In the third stage of listening comprehension, the utilization stage, listeners utilize the mental representation of the sentence by showing such actions as storing the sentence in their minds, answering a question, obeying an order and so on. Anderson (2000) also emphasizes that these stages are mostly obligatorily in an order, but they mostly overlap, too. For example, a listener may still be making inferences about an earlier part of a sentence and simultaneously

perceiving a later part of it. Therefore, it can be argued that a strict order between these stages of comprehension does not exist.

Field (2009) also provides a process view of listening, which focuses on the integrative and interactive nature of listening. Field's (2009) model distinguishes between two important stages in listening comprehension: "decoding and meaning building" (p.125). He argues that decoding refers to making sense of the speech signal in order to recognize words and then identifying a grammatical pattern based on the words that are recognized. Meaning building, on the other hand, has two main aims. Firstly, the listener tries to understand the meaning of what the speaker says at word and sentence levels based on the context and the situation. Secondly, the listener evaluates the information in terms of its importance and forms an overall idea of the message with the relevant information.

Drawing upon Cutler and Clifton (1999), Anderson (2000), and Field (2009), Field (2013, pp.95-6) proposes a comprehensive model of the listening process, which consists of five levels of processing:

1. input decoding
2. lexical search
3. parsing
4. meaning construction
5. discourse construction

In the following pages, Field's cognitive processing framework of listening comprehension is explained with further details since it provides a clear and comprehensive picture of the listening comprehension process and also the listening construct in this study is defined according to this framework. Field divides the five levels of listening comprehension into lower-level and higher-level processes; therefore, the processes are examined according to these categories.

2.3.2.1 Lower-level processes

The lower-level processes in this listening framework include the first three levels of processing in the model; namely, input decoding, lexical search and parsing. These are widely researched levels as they explain the fundamental processes explaining L1 speech recognition and have suggestions for L2 listening comprehension as well.

The first step in comprehension of speech is input decoding. After the perception of speech-like sounds, the listener attempts to match them against the phonological system of the language and forms representations in his/her mind. However, Field (2013) also states that variation can happen at the level of phonemes during natural speech. Due to this, it can be difficult to explain how analysis works at this level. Therefore, it has been argued that analysis at syllable or word level is more important than at phoneme level (Field, 2013).

After the recognition of sounds, lexical search is required. Field (2013) states that at this level the listener maps the sequences of sounds to the words in their lexicon. It also means that a number of possible words are identified with the help of the evidence available and the sound signal is matched to one of these words as the listener receives more evidence. Field (2013) also draws on other theories attempting to explain how words are recognized. One of them is cohort theory, which entails that the recognition of words happen during a process of matching the sounds to words in time (Marslen-Wilson, 1973, in Field, 2013). It is suggested that listening is an event that happens online and that the listener tries to decode the linguistic input as quickly as a syllable instead of waiting for the sentence to come to an end (Field, 2013). The information provided at the beginning of the words, perhaps the first syllable, activates some words, which form the cohort (Rost, 2013). The recognition of a word by its first clues leads to the opening of a cohort of possible matches

(Field, 2013) and all the words that begin with the first sound of the word are activated (Rost, 2013). As the listener receives more input, the options in the cohort are eliminated until a match is found (Rost, 2013; Field, 2013).

However, Marslen-Wilson's assumptions were challenged by the findings of Grosjean (1985) in experiments where gating method was used and listeners were exposed to larger chunks of a sentence and asked to report what they thought they heard. As a result of these experiments, it was revealed that not all words could be identified correctly before their acoustic offset. It was also suggested that word recognition is not strictly a word-by-word or sequential process; rather, it is retroactive, and listeners go back and forth to adjust their earlier analyses of sounds (Grosjean, 1985). However, according to cohort theory, only the acoustic-phonetic information seems to cause activation for the correct word in candidates (Rost, 2013). Therefore, Rost claims that cohort theory fails to account for a complete listening comprehension process and it provides a model which suggests that linear cues from the speech are used to understand the spoken message.

Another theory explaining word recognition is logogen theory by Morton (1969, in Rost, 2013). Logogen theory provides insight into the interactive process of speech decoding and lexical access. 'Logogens' are sensing devices that a listener has and that represent the words in the mental lexicon of the listener. A logogen is composed of all the information related to a word and checks auditory or visual/graphic information for a match. In case of an encounter with the information that a logogen carries, the logogen is activated and the information related to the word is made available to the 'response system' of the listener. After the recognition of the word, lexical access is considered to be automatic (Rost, 2013).

Another model of lexical search was proposed by McQueen (2007). The words in the listener's lexicon compete with each other as a result of the recognition of the signal. The amount of activation that these words receive based on how similar they are to the speech signal determines the recognition of the correct word. This premise outlines an interactive view of lexical retrieval similar to logogen theory.

An issue of discussion in listening comprehension at word recognition level is that determining the word boundaries can pose problems to the listener, as regular pauses between words do not exist in real connected speech unlike written language (Field, 2013). Cutler (1997, in Field, 2013) points out that lexical stress, syllabicity and vowel harmony are features that help listeners determine where a word begins and ends. Prosody can play a role in understanding word boundaries; for example, by signaling the content words with more stress and function words with less stress. Moreover, Field (2013) argues that words can vary during speech. The way words are pronounced can differ for three main reasons; the importance of the word, the complexity of pronunciation and the formality of the speech. These may impact on the word recognition process and hinder comprehension.

During the mapping of the sounds to the mental lexicon, the listener can also make use of clues such as the frequency of the words relative to each other. Field (2013) points out that in the sentence "We never found" the listener is more likely to think of the word "found" as the past simple form of the verb "find" instead of the present simple form of the verb "found" because of the relatively higher frequency of the verb "find" and its past simple form. Another process that helps the recognition of words is called spreading activation (Field, 2013). The complex lexical networks in our brain make it easier for us to recognize words which are associated with previously mentioned ones. For example, after hearing the word "school", it will be

quicker for a listener to recognize such words as teacher, student, homework, etc.

Field (2013) also emphasizes that at lexical search level the listener not only matches the signal to possible words but also to the meanings of these words. The meanings and forms of these words can only be finalized after hearing the whole intonation and syntactic structure of the utterance.

Another lower-level process in Field's model of listening is parsing. Field (2013) argues that parsing is an event that occurs online at the same time as the production of the utterance. During listening the listener makes predictions as to the upcoming structures and confirms or refutes his/her predictions based on the new evidence. After assigning syntactic structures to the words, the listener may also need to choose from the possible meanings of a word in the utterance. Once parsing has finished, listeners convert the things that they have heard into a proposition, "an abstract representation of a single idea", as maintaining a lot of words in the mind can impose cognitive load on the mind unnecessarily (Field, 2009). The propositional meaning refers to the literal meaning of the clause, which does not include the context, or the background knowledge of the listener. The propositional meaning replaces the linguistic form that has contained the message up to that point. The listener stores the idea of the clause, but they may not remember the exact words used to convey the message. This indicates the output of the parsing process.

2.3.2.2 Higher-Level Processes

The last two levels of processing in Field's model of listening comprehension are meaning and discourse construction, which are considered as higher-level processes. The meaning attached to the utterance during parsing, i.e. propositional meaning, is abstract and literal without regard to the context in which listening occurs. The

listener needs to utilize contextual information, topic knowledge, world knowledge, etc. in order to understand the relationship between the proposition and the immediate situation (Field, 2009). When the listener combines these types of knowledge with the propositional meaning, they form meaning construction, or an enriched meaning of the proposition (Field, 2009). At meaning construction level, the listener reaches the implied meaning of the utterance with the help of contextual clues or knowledge of the listener. Schematic knowledge plays an essential role in making inferences and arriving at the speaker's intended meaning. Field (2009) describes schema as a complex knowledge structure in a person's mind including everything that s/he knows about and associates with a certain concept. Schematic information can help listeners in two main aspects. Firstly, they can form presuppositions on the listening text by using their schematic information. Secondly, because of the shared knowledge the listener and the speaker have, many things can be left unsaid; as a result, the listener can fill in the necessary information thanks to schematic information and make the necessary inferences (Field, 2009). In addition to the role of schematic knowledge, there are other sources such as pragmatic, contextual, semantic and inferential knowledge sources that the listener can draw upon to grasp the implied or intended meaning of the sentence beyond the literal, or abstract, one (Field, 2013). All of these combined assist the listener to achieve successful comprehension of the spoken message at both propositional and meaning representation levels.

After forming an enriched meaning in the minds, the meaning representation needs to become a part of the listener's memory of what has been said so far. In this way, the listener creates a discourse representation in their minds with all the messages combined, analyzed, synthesized, etc. (Field, 2009). Field (2013) maintains

that during listening, the listener continuously judges the information that he/she has received up to that point and relates it to his/her previous knowledge. As a result, he/she forms a recall of what has been said so far. While doing this, the listener utilizes some processes to construct a wider meaning of the listening input. Selection, integration, self-monitoring, and structure building are the processes offered by Field (2013). Selection refers to determining whether a piece of information is related to the discourse as a whole or the speaker's intentions. Integration entails the addition of new pieces of meaning, especially those bearing conceptual links such as result, contrast, etc., to the discourse representation in order to link the upcoming information to the previous information. Self-monitoring requires the listener to check their understanding of the links and consistency between the old and new meaning units. Structure building enables the listener to make a hierarchy between the meaning items according to their relative importance in the discourse. The listener can form major and subordinate points about the message and store the information in such a way. By using these processes, the listener attempts to form an overall idea of the spoken message.

As mentioned in the lower-level processes, both higher-level processes are not necessarily sequential, either. This five-level processing framework is regarded as the main listening model for this study in terms of theory-based and context validity claims and references to it will be made frequently in the upcoming chapters to discuss these validity types.

2.3.2.3 CEFR descriptors for the listening skill

In addition to Field's (2013) framework for the listening skill, specifications for language abilities offered by the Common European Framework of reference can be

utilized in order to define language constructs and determine the target sub-skills of a construct to be assessed. The CEFR offers a common ground for the development of syllabuses for language courses, curriculum, examinations, course materials, etc. It is developed by the Council of Europe (2001) in an attempt to describe comprehensively what language learners need to learn in order to be able to communicate through a language and what knowledge and skills they need to develop to be able to do so. Using the CEFR descriptors enables course designers, teachers, test developers, etc. to provide a sound basis for their works. To this end, the CEFR provides descriptors of different language skills at six different proficiency levels ranging from A1 to C2 demonstrating the progress of the learners at each stage of learning. In the CEFR, descriptors of the four language skills (speaking, listening, reading and writing) in terms of production, reception and interaction as well as general competences and communicative language competences are stated separately for the six proficiency levels. The CEFR descriptors for the overall listening ability are shown in Appendix A. These descriptors guide test developers for the alignment of their tests and test tasks to the different proficiency levels.

2.3.3 Context validity

In addition to the validity concerns regarding the construct and the cognitive processes related to that construct, test developers also need to investigate context validity. As Weir states, listening, or other skills for that matter, “does not take place in a vacuum” (2005, p.19). Therefore, there is a need to describe the context in which language operations are performed, too. The traditional term “content validity” is referred to as “context validity” in Weir’s socio-cognitive framework since it reflects the social aspect of language use better. He explains context validity as follows:

Context validity is concerned with 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. This coverage relates to linguistic and interlocutor demands made by the task(s) as well as the conditions under which the task is performed arising from both the task itself and its administrative setting. (Weir, 2005, p.19)

As it can be seen from this description, context validity requires test tasks to be as similar to real-life conditions as possible. Achieving context validity ensures that the test tasks represent the target language use sufficiently in terms of the target task setting, administration setting, and task demands in terms of linguistic features of the tasks and interlocutors, or speakers. According to Weir (2005), the contextual features of the tasks should be explored in terms of their similarity with the TLU domain or, in other words, their authenticity. This necessity has been underlined by several other researchers who caution that the issue of authenticity in test design should be given careful consideration. Bachman and Palmer (1996) define authenticity as “the degree of correspondence of the characteristics of a given language test task to the characteristics of a TLU task” (p.23). They mention two kinds of authenticity; interactional and situational authenticity. The former is related to the authenticity of the interaction between test takers and test tasks in terms of the cognitive processes they need to employ to complete the task. On the other hand, the situational authenticity of a test task is dependent on the relationship between its test method characteristics and the features of the TLU domain. The tasks selected for a test should be representative of real-world listening events. Weir (2005) states that situational authenticity is primarily related to context validity while interactional authenticity is linked with theory-based validity. Therefore, achieving situational authenticity contributes greatly to the context validity of a test.

Weir (2005) argues that in test design, various elements concerning task setting, administration setting, and task demands in terms of linguistic characteristics

and speakers should be taken into consideration in order to form a theoretically sound basis for the choices made regarding contextual features of the test tasks. Investigation of these elements and their relation to the test tasks will provide test developers with evidence to validate the inferences they would like to make about test takers based on their test performances. Therefore, the essential components of context validity as is shown in Figure 2 are examined below in more details.

SETTING: TASK • Purpose • Response format • Known criteria • Weighting • Order of items • Time constraints	DEMANDS: TASK • Linguistic: ✓ Discourse mode ✓ Channel ✓ Length ✓ Nature of information ✓ Content knowledge ✓ Lexical ✓ Structural ✓ Functional • Interlocutor ✓ Speech rate ✓ Variety of accent ✓ Acquaintanceship ✓ Number of speakers ✓ Gender
SETTING: ADMINISTRATION • Physical conditions • Uniformity of administration • Security	

Figure 2. Aspects of context validity for listening (Weir, 2005, p. 45)

2.3.3.1 Task setting

Task setting refers to the setting under which test administration takes place and the components of task rubric (instructions) that should normally be provided to test takers. As shown in Figure 2, task setting is composed of six important parameters and each is discussed below respectively.

One crucial aspect of task setting is the purpose as outlined by Weir (2005). The purpose of the task is related with authenticity, which is discussed above in section 2.3.3. Widdowson (1978, in Lynch, 2009) argues that what is essential is whether a listening text is genuine; i.e. it resembles the kind of language that can be

used in a similar real life situation. Bachman and Palmer (1996) discuss that what lies at the heart of authenticity is the correspondence between the features of the TLU tasks and those of the test task. The degree of authenticity can tell us the extent to which the interpretations based on the scores can be generalized to the TLU domain. Taking both arguments into account, we can suggest that authenticity of the input (the text) and the output (the task) has implications for test development. Task purpose also has implications for task type. It determines the purpose for which test takers are listening to a text. Field (2009) suggests a model of task purpose for listening based on Urquhart and Weir's (1998, in Field, 2009) classification of task types for reading comprehension. Instead of the expeditious reading category, he included four levels of attention allocated to listening in his classification. Figure 3 shows Field's (2009) modified version of task purposes for listening, which is consisted of global and local listening skills and four levels of attentional focus. Field (2009) argues that task types that require low level of attention are more suitable for listeners at low levels of proficiency. When they become more competent at lexical recognition, they can be expected to comprehend various facts, messages, and relationships in the text. Similarly, when they do not need to focus on lower-level processes, they can allocate their resources such as memory for meaning and discourse construction¹. One final aspect of task purpose is related with the rubric or the instructions that we provide to test takers and different types of task types should be taken into consideration while preparing the rubrics (Elliott & Wilson, 2013).

¹As Weir (2005) states, context validity and cognitive validity have an interdependent relationship. Especially task types in a test have to be discussed with reference to the cognitive processes that they trigger and the cognitive load they impose upon test takers; therefore, referring to cognitive validity while discussing task types should not be considered irrelevant in the discussion of context validity.

	Global	Local
Shallow attentional focus	<i>Skimming</i> (listening generally) to establish discourse topic and main ideas. <i>'What is it about?'</i> e.g. TV channel hopping, TV advertisements, eavesdropping <i>Phatic communion</i> <i>'What are the speaker's intentions?'</i> e.g. greetings	<i>Unfocused scanning</i> to locate information relevant to the listener. <i>'Does the speaker mention anything of interest to me?'</i> e.g. news headlines
Medium attentional focus	<i>Listening for plot; listening to commentary</i> <i>'What happened next?'</i> e.g. film/TV drama, TV/radio interview <i>Conversational listening</i> <i>'What is the speaker's message?'</i> e.g. everyday chat <i>Information exchange</i> <i>'How much do I need to know?'</i> e.g. tour guide	<i>Focused scanning</i> to locate one area of information needed by the listener. <i>'When will the speaker mention X?'</i> e.g. airport announcement, weather forecast <i>Search listening</i> to locate and understand information relevant to predetermined needs. <i>'What is the answer to these questions?'</i> e.g. hotel/travel information <i>Message listening</i> <i>'How many details do I need to retain?'</i> e.g. answerphone
Deep attentional focus	<i>Close listening</i> to establish the speaker's main points and to trace connections between them. <i>'What is important?'</i> e.g. lecture listening	<i>Close listening</i> to record in depth the speaker's main points and supporting detail. <i>'I assume that everything is relevant.'</i> e.g. negotiation
Very deep attentional focus	<i>Listening to check critical facts</i> <i>'Is this consistent?'</i> e.g. witness evidence	<i>Listening to vital instructions</i> <i>'I assume that everything is important.'</i> e.g. street directions <i>Listening to the form of words</i> <i>'What precisely did he say?'</i> e.g. listening to quote somebody

Figure 3. Types of listening as determined by listener's goals (Field, 2009, p.66)

Another important facet of task setting is response format. Elliott and Wilson (2013) argue that each response method has their benefits and drawbacks; therefore, including a variety of different response formats in tests can enable test writers to eliminate the disadvantages that may arise from test format. In addition, different response formats may be more suitable for different cognitive processes and text types. Furthermore, different groups of candidates may be at a disadvantage if the expected response format is unfamiliar to them. Thus, the choice of response format is an essential issue in test development.

Elliott and Wilson (2013) list some important elements to take into consideration before choosing the response method (2013, pp.162-3). This list of important considerations is definitely not a comprehensive one; however, it raises some important concerns about the response format to be chosen. These considerations can be listed as follows.

- The rubrics should be clear and unambiguous, and there should not exist any correct answers other than those indicated in the key.
- The item stems should not be more difficult than the recorded text linguistically, the opposite of which may cause construct-irrelevant variance.
- Whether test takers are allowed to take notes during listening should be specified.
- The memory load imposed on test takers by the response method should be taken into account.

In addition to the parameters mentioned above, the types of response formats included in a test should also be given careful thought before developing a test. There are a variety of response formats that can be used to assess listening. In *Examining Reading*, Khalifa and Weir (2009) divided response formats into two categories as selected response formats and constructed response formats. In the former, test takers are supposed to choose from the available options whereas in the latter they are supposed to produce the answer themselves by writing mostly. These response formats can be used for listening tests as well. Under these categories, two response formats are used widely in the assessment of listening: multiple-choice questions (MCQs) and short-answer questions. The pros and cons of using these response formats are discussed below thoroughly.

MCQs offer test writers many benefits and the ease of marking, therefore reliability. Firstly, MCQs can target different levels of processing such as lexical search, parsing or meaning construction. They also provide test writers with flexibility when listening texts cannot be recorded again, as the options can be modified easily. Besides, writing as a construct-irrelevant variance is eliminated from the tasks since test takers only mark the correct response.

On the other hand, several concerns have also been raised for this response format, especially in terms of the interaction between context and cognitive validity (Field, 2013). In a multiple-choice task, test takers need to carry out more than one process at the same time. They need to listen for the correct answer, keep the options in mind, evaluate the options according to the incoming information, confirm the correct option and also generally disconfirm the wrong options. Doing all of these processes at the same time is a very complex operation and imposes a heavy cognitive load on test takers. In addition, the manner listening event occurs in this test environment does not resemble a non-test one, which is a challenge to cognitive validity. Furthermore, the two different modalities used in the test, spoken and written input sources, make things even more complicated from a cognitive perspective. Another issue related to this test format is that test takers' familiarity may in fact prove to be a disadvantage when test-wiseness strategies are employed. Test takers may be trained to look for loopholes in the test, which is by-product of the test method, not the listening construct. The nature of the task can sometimes encourage test takers to use test-wiseness strategies and choose the correct answer by eliminating the distractors rather than identifying the correct answer, which raises questions related to cognitive validity (Field, 2013). Since no response format is perfect, test writers should do their best to eliminate these drawbacks of multiple-choice tests if they are to use them.

In addition to the concerns mentioned above, another consideration in preparing a multiple-choice test is the number of options. According to a study conducted by Rodriguez (2005, in Elliott & Wilson, 2013), three options may work nearly as well as four or more options for MCQs in terms of item difficulty and discrimination. This could help decrease the memory and reading load on test takers.

However, in some other studies (Moreno, Martinez & Muñiz, 2006; Boroughs, 2003; in Elliott & Wilson, 2013) four-option MCQs are deemed more appropriate in terms of item difficulty and discrimination, especially for high proficiency levels.

The other response format, short-answer questions differ from MCQs in that they do not give as many clues as MCQs to test takers and do not interfere with the cognitive processes as much. They resemble real-life listening events more as test takers are not guided by any options and do not need to deal with options. In addition, guessing factor is significantly decreased, though test takers may make some predictions as to the correct answer because of the linguistics clues in the items (Elliott & Wilson, 2013).

However, an important drawback of short-answer questions is the evaluation of spelling. Spelling accuracy is not a part of the listening construct, so whether to accept spelling mistakes or not will have implications with respect to cognitive validity. There are three strategies to follow while marking constructed responses (Elliott & Wilson, 2013, p.168).

1. Accepting all plausible phonetic misspellings of a word.
2. Accepting a limited, prescribed range of misspellings of a word and not others.
3. Accepting only the correct spellings of a word.

The first two options are likely to cause problems regarding objectivity, as different markers can accept different misspellings as correct or incorrect. Especially in large-scale tests, the problem could get bigger.

Another issue should also be taken into account about marking constructed response formats is the clarity of the key. Synonyms of some words could sometimes be accepted; however, when the key includes a broad range of correct answers, it may look complicated and make marking challenging. In addition, if the correct answers require certain grammatical structures such as tenses or affixes, this may

also cause confusion and difficulty in marking. Therefore, constructing items with answers that do not demand any grammatical modifications would help avoid problems.

All of the discussion provided above regarding response format demonstrates that choosing a specific response format and task type in a test requires careful consideration and the benefits and drawbacks of each response format should be evaluated meticulously if we would like to achieve context and cognitive validity.

Another consideration in terms of task setting is known criteria, i.e. informing test takers about the way they will be assessed. Test takers need to know how their scores will be marked and if there are any criteria for correct answers. Weir (2005) discusses that knowledge of criteria may affect test takers' cognitive processes and lead them to use different strategies. Therefore, information regarding criteria should be specified in the rubrics or any other relevant sections in the test.

Weighting is also pointed out by Weir (2005) as related to task setting. Weighting means that different marks are given to different items in the same task or to different tasks in the same test. This implies that tasks or items with higher marks are more important than the others. If weighting of items or tasks is done in a test, test takers need to be informed about this situation as they may want to manage their resources such as memory and time depending on the importance of the items and the tasks (Elliott & Wilson, 2013).

Order of items in a listening test is another essential component of task setting. Listening is an online activity and listeners cannot go back and forth while listening. During listening, test takers build meaning and discourse representations in their minds with the available incoming information. Therefore, listening takes place in a cumulative way. Due to this nature of listening, the order of items in a listening

test is crucial. When items are presented in a different order than the text, it will cause extra cognitive load on test takers and interfere with the cognitive processing during comprehension and meaning construction. Thus, it is of great importance that items are ordered according to the recorded text (Weir, 2005).

Time constraints also need to be considered in terms of task setting during the development of a listening text. Time constraint refers to how much time will be allocated to test takers between two listening tasks and between two items in a task and how many times test takers will hear the listening text. Elliott and Wilson (2013) suggest that the amount of time between items and tasks should suffice for test takers to answer one question or complete the task and get ready for the following one in order not to cause construct irrelevant-variance. Furthermore, the times of playing the recording should be determined as well. Listeners can ask for clarifications or repetition and hear the information again; however, they may not hear the exact same words or different cognitive processes could be employed at the second time of hearing (Field, 2013). A counter-argument against single play is put forward by Murray (2007, in Field, 2013). It is stated that as technological advances have increased, people's opportunities to listen to things such as videos, radio podcasts, TV programs, etc. as much as they want have increased, too. Therefore, double-play is considered to be a part of real-life and should be applied in test situations. Taking both views into account, we can state that single or double play issue is still to be settled in listening assessment.

2.3.3.2 Administration setting

Administration setting is concerned with the conditions under which the test is administered and the efficiency and reliability of the administration. Weir (2005)

states the significant components of administration setting that must be carefully designed are physical conditions, uniformity of administration and security. Elliott and Wilson (2013) maintain that there must be clear procedures as to the administration of the exam. In addition, examination staff should be informed about a number of important topics such as the test, the instructions, test equipment, policy about latecomers and cheating, etc. Classroom conditions, e.g. chairs, desks, heat, clocks, etc., must also be appropriate for a test setting, and the test equipment; for example, loudspeakers, laptop computers, etc., should be checked in advance and any possible problems should be mitigated. The quality of the recordings, the sound system in the exam venue, and the quality of the exam papers should be ensured prior to the test administration. Precautions related to the security of the exam papers and the exam venue should also be taken in order to create a safe testing environment.

2.3.3.3 Task demands (Linguistic)

Task demands are divided into two categories in Weir's (2005) framework: Linguistic and interlocutor demands. Firstly, the linguistic demands of the tasks will be discussed according to the components as shown in Figure 2 in section 2.3.3.

Discourse mode entails that texts chosen for listening tests should be representative enough of the texts that test takers may deal with in the TLU domain. The overall text purpose specifies what the discourse aims and what test takers are led to achieve in the test (Elliott & Wilson, 2013). Kinneavy (1969) focuses on discourse modes that can be related to text purposes and the intention of the speaker. As it can be seen from Figure 4, different genres can fall within the same discourse

mode. During the development of a listening test, the tasks and their discourse modes should be considered to be able to achieve better context validity.

Referential	<i>Informative</i>	e.g. news, articles, reports, summaries
	<i>Scientific</i>	e.g. proving a point by arguing from accepted premises, proving a point by generalising from particulars, a combination of both
	<i>Exploratory</i>	e.g. dialogues, seminars, a tentative definition of..., proposing a solution to problems, diagnosis
Persuasive		e.g. advertising, political speeches, religious sermons, legal oratory, editorials
Literary		e.g. short story, lyric, short narrative, limerick, ballad, folk song, drama, TV show, movie, joke
Expressive	<i>of individual</i>	e.g. conversation, journals, diaries, gripe sessions, prayer
	<i>of society</i>	e.g. minority of protests, manifestoes, declarations of independence, contracts, constitutions of clubs, myth, utopia plans, religious credos

Figure 4. A system of text purposes (Adapted from Kinneavy 1969, p.302)

Channel of presentation is another facet of linguistic demands of the tasks. It refers to the way the input is provided to test takers. In listening tests, there are generally two different channels of presentation. The recorded text is in spoken - audio- form and the items are in written form. In a listening test, the necessary information needs to be delivered unambiguously and cautiously via both channels of presentation. The quality of the audios should be checked carefully before the administration of the test since it provides the main source of information in a listening test. In addition, the written information, i.e. the components of the tasks such as the items and the instructions, needs to be legible and sensible to the test

takers so that construct irrelevant variance does not occur during the administration of the test.

Text length is another consideration in terms of linguistic demands of the test tasks. It is concerned with the length of listening texts as a recording and also the number of words in listening texts. These should be carefully controlled while preparing listening tests and their suitability for different proficiency levels should be checked beforehand in order to ensure context validity.

The nature of information included in a listening text also affects test takers' performances. The level of concreteness and abstractness can change depending on the target proficiency levels and can impose different cognitive load on the test takers; therefore, the use of concrete and abstract words should be monitored closely in listening texts.

Texts can place certain demands on the listener in terms of background and subject knowledge, too. These two types of knowledge should be taken into consideration in test design, as variance in scores due to them causes construct irrelevant variance unless such knowledge is specifically required in the test. Therefore, test writers need to design tests in a way that would not put those who lack knowledge about the topic at a disadvantage.

Another facet of linguistic demands, lexical resources used in listening texts impact test performance considerably. One of the most basic processes in listening comprehension is lexical search as mentioned before. Consequently, the amount of vocabulary that listeners have in their repertoire influences the understanding of the listening text (Elliott & Wilson, 2013). In a study carried out by Stæhr (2009), the role of vocabulary knowledge in listening comprehension was investigated in an EFL classroom with 115 advanced Danish learners. The study revealed that listening

comprehension is significantly correlated with vocabulary knowledge and it accounts for half of the variance in listening test scores. Moreover, it was suggested that a lexical coverage of 98% is required in order to be able to handle the listening texts. This coverage was also reported to be consistent with findings from reading research. However, Van Zeeland and Schmidt (2012) argue that at least for everyday narratives a lexical coverage of 90 to 95 percent seems sufficient for successful comprehension.

Elliott and Wilson (2013) also mention that knowing the word is not sufficient on its own, as listeners also need to recognize and identify the word and match it against the possible words in their mind. Lexical search in listening differs from reading in that listeners do not have a written form in front of them and they need to cope with other factors such as unclear word boundaries, different forms of pronunciation and word-linking features such as elision and assimilation. In addition to dealing with these factors, there are other issues such as polysemy and homophony, which require a wider coverage of knowledge of the words. Furthermore, as well as the amount of words that are known, the amount of knowledge that a listener possesses about these words also contributes to listening comprehension. Nation (1990, in Elliott & Wilson, 2013, pp.218-9) suggests a multi-dimensional taxonomy of depth of knowledge of a word:

1. Spoken form of the word.
2. Written form of the word.
3. Grammatical behaviour of the word.
4. Collocational behaviour of the word.
5. Frequency of the word.
6. Stylistic register constraints of the word.
7. Conceptual meaning of the word.
8. Associations the word has with other related words.

The taxonomy of word knowledge by Nation (1990) demonstrates that knowing only the meaning of the word may not be sufficient for a complete understanding of an

utterance and having more comprehensive knowledge about a word can facilitate comprehension.

In addition to the lexical resources, the grammatical resources that test takers need to employ in a listening test need to be taken into account to control the linguistic demands of the tasks. Syntactic parsing is one of the initial steps of listening comprehension according to Field (2013); therefore, the grammatical structures in a text should be evaluated carefully. They should reflect the kind of grammatical structures used in the TLU domain so that test takers' performances can be generalizable to real-life performance. At this point, an important issue emerges: the differences between spoken and written language. Spoken language has its own features such as inconsistencies, false starts, word stress, intonation, etc. This means that while preparing listening texts, test writers also need to consider the features of spoken language.

The grammatical complexity of a listening text inevitably affects comprehension, especially when test takers are not familiar with the structures. It is argued that listeners at lower proficiency levels depend more on the grammatical structures for comprehension than listeners at higher proficiency levels, who automatically process grammatical structures and use semantic cues for comprehension (Conrad, 1985, in Elliott & Wilson, 2013). Therefore, especially for lower-level tests, the control of grammatical complexity is essential and test writers need to take into consideration the grammatical requirements of the items carefully. In the Reference Level Descriptors prepared based on the CEFR scales (Breakthrough (Trim, 2009), Waystage 1990 & Threshold 1990 (van Ek and Trim, 1991) and Vantage (van Ek and Trim, 2001)) the minimum language requirements for the relevant proficiency levels are defined and grammatical structures which can

be used at certain proficiency levels are suggested. These suggestions provide test developers, course designers, materials writers, etc. with insight into the expected developments of language learners.

Besides the lexical and grammatical demands of the tasks, the functional resources required by the test tasks need to be carefully monitored, too, in order to be able to align the tasks across different proficiency levels. In the Reference Level Descriptors, some broad categories of language functions are also stated. These categories from Vantage (2001) are listed below:

- Imparting and seeking factual information
- Expressing and finding out attitudes
- Deciding on and managing courses of action: suasion
- Socialising
- Structuring discourse
- Assuring and repairing communication repair

Under these broad categories, there are a wide variety of specific language functions specified and exemplified according to different proficiency levels. These examples and specifications can be taken as a reference for test development and help test takers develop standardized tests which can be shown to align with the requirements of the target proficiency levels.

2.3.3.4 Task demands (Interlocutor)

Task demands in terms of interlocutors are related with the features of the speakers such as speech rate, number and gender of the speakers, accent, acquaintanceship, etc. Speech rate may affect comprehension, especially at lower proficiency levels (Elliott & Wilson, 2013). Fast speech may impede understanding of lower-level learners, as their listening skills are not automatized yet. Therefore, it is important to take into account the speed at which the texts are recorded and the differences

between the speech rates of listening texts at different proficiency levels should be considered carefully.

In terms of accent, there has been a growing discussion of the inclusion of different accents in listening tests for English language (Field, 2013). The widespread use of English as a lingua franca in the world has increased the frequency of different accents in listening tests. Therefore, whether different accents need to be included in a test should be carefully thought in line with the requirements of the TLU domain.

Weir (2005) states that whether listeners are familiar with the voice of the person they are listening to may have an influence on their comprehension. The degree of acquaintanceship can be controlled easily for low-stakes tests which are given in a classroom, since the teachers can read the listening text; however, it is almost impossible to achieve a high level of familiarity in a standardized and high-stakes test. Therefore, the best we can do as test designers can be to choose interlocutors with a clear and comprehensible voice.

Elliott and Wilson (2013) state that the number of speakers to be included in the texts should also be determined as a part of interlocutor demands. He states that the number of speakers is mainly determined by the TLU domain and the target real-life situation. If a group discussion is added as a text, then a number of speakers may be needed whereas for a doctor-patient relationship as in A1 level task two speakers seem to be sufficient. Similarly, for announcements, one speaker would suffice.

Genders of the speakers also need to be considered in order to avoid cultural bias. Furthermore, when the listening texts require more than one speaker, the gender of the speakers are very important as it can make it easy or difficult for test takers to

differentiate between the different voices. Consequently, the genders of the speakers should cautiously observed in listening texts.

In conclusion, the important parameters of context validity as outlined by Weir (2005) should be taken into consideration before the administration of a listening test in order to demonstrate that the requirements of the TLU domain are met across different proficiency levels during test development.

2.3.4 Scoring validity

Weir (2005) states that scoring validity is another term that covers all sorts of reliability. Scoring validity pertains to “the degree to which examination marks are free from errors of measurement and therefore the extent to which they can be depended on for making decisions about the candidate” (Weir, 2005, p.23). Bachman and Palmer (1996) simply define it as “consistency of measurement”. Geranpayeh (2013) emphasizes that scoring validity is a significant aspect of test validation since it directly affects the scores obtained from the test and the decisions we make based on those scores; therefore, problems of inconsistency, unsystematicity or test administration can decrease the validity of the test and can lead to the involvement of construct-irrelevant variance in the testing process.

Language tests can be variable, or inconsistent, due to a number of reasons related to test taker characteristics such as test-taking strategies, illnesses, boredom, anxiety, fatigue, etc., and to some test-related factors such as poorly-designed tasks, number of tasks and task types, ambiguous instructions, poor scoring methods, etc. even though the skills tested remain the same (Douglas, 2010). Identifying possible sources of error in a language test is critical in test design and the effects of these errors should be minimized since they will inevitably impact test performance and

test scores, and thereby the validity of the interpretation of the scores (Bachman, 1990). Bachman (1990) also argues that reliability concerns should be investigated through logical analysis and empirical research. This means that we, as test developers, need to detect sources of error and then carry out statistical analyses to “estimate the magnitude of their effects on test scores” (Bachman, 1990, p.161).

In addition to Weir’s view of reliability as part of a test’s overall validity, other researchers also discuss the relationship between validity and reliability from a slightly different perspective. Reliability, as a statistical inquiry, is a prerequisite condition for validity since a test score which is unreliable cannot be valid (Bachman, 1990). It is also stated that these two crucial aspects of tests should be seen as complementing each other rather than as two distinct concepts (Bachman, 1990; Geranpayeh, 2013). Reliability is concerned with data quality, and validity is related to the meaningfulness and appropriateness of the inferences we make about test scores (Geranpayeh, 2013). Similarly, according to Bachman (1990), reliability and validity are associated with two complementary aims in test design and development: “(1) to minimize the effects of measurement error [reliability], and (2) to maximize the effects of the language abilities we want to measure [validity]” (p. 161). Therefore, investigation of both validity and reliability provides evidence related to different aspects of tests.

McNamara (2000) maintains that scoring validity looks into how well the process of assessment is by investigating scores, and the data we depend on come in two ways. Firstly, we mark performances in the assessment by assigning numbers or scores to them and obtain a set of scores from the test takers’ performances. Secondly, we conduct statistical analysis on these scores and investigate the properties such as item facility, item discrimination or reliability for these scores. By

doing these, we attempt to make sure that we can make meaningful and fair decisions about the test takers (McNamara, 2000).

Moreover, in test design, as well as the test content, we need to consider the test method, which indicates how the test takers will respond to the test items and how the raters will rate or score their answers. Establishing a rating procedure is an essential component of the assessment process because we would like to account for the ratings of test takers and rate their performances in a standardized and systematic way (McNamara, 2000).

McNamara (2000) argues that there are some quality control procedures that we need to follow to ensure the meaningfulness and fairness of assessment. Item analysis is an essential element of this procedure, which helps us evaluate the effectiveness of the test items. It is a part of trialing as well as post-exam data collection process. Bachman (2004) discusses that classical item analysis (CIA) and item response theory (IRT) are two procedures that can be used to understand the characteristics of individual test tasks and items better. In this study, CIA will be employed due to the number of the participants in the study and its common use in test analyses. The main aspects of CIA that can be examined are measures of central tendency, item characteristics and reliability, on which the current study focuses for the investigation of scoring validity. The statistical analyses conducted for the test scores in the present study are discussed thoroughly in Chapter 4 section 4.4 under the investigation of research question 3.

2.4 Test development

Test development is the process during which we create and administer tests.

Bachman and Palmer (1996) maintain that the type of test we are preparing and our

purpose for using it will determine how much effort we will put into it. For example, preparing an informal low-stakes test may not require much time and effort, whereas a national high-stakes test may require a whole test development team, as it may be trialed and revised several times. However, this should not mean that we sacrifice the concerns for the validity of our interpretations in low-stakes tests. Both for high and low-stakes tests, we need to make sure that the components of test usefulness are taken into account. Therefore, no matter what type of test we are preparing, we need to think thoroughly and make a lot of preparations. Bachman and Palmer (1996, p.9) state their approach to test development and use with these basic principles:

1. the need for correspondence between language test performance and language use
2. a clear and explicit definition of the qualities of test usefulness

Bachman and Palmer (1996) suggest that once we decide on the TLU domain to which we would like to generalize the performances on our test, we also need to decide on the test characteristics such as the task, the topic, the test items, the skills to test, etc. We also need to take into consideration the characteristics of the test takers – their prior knowledge, affective schemata, and test taking strategies. These issues mentioned by Bachman and Palmer (1996) are parallel to the components of Weir's test validation framework emphasizing crucial points about test development.

McNamara (2000), stating that creating a new test involves “a design stage, a construction stage and a try-out stage” before we actually administer it, also discusses that although these steps seem linear, test developers always need to make some revisions at different stages; thus, test development is in fact a cyclical process. Similarly, Bachman and Palmer (2010) maintain that assessment development and use includes many activities and these activities can be conceptually ordered as follows: 1) Initial Planning, 2) Design, 3) Operationalization, 4) Trialing, and 5)

Assessment Use. However, they also point out that although these activities are ordered in a way, test development is in fact an iterative process, which means that test developers may need to revise and make some changes regarding other stages of test development. Therefore, this should not be seen as a strict order.

Douglas (2010) suggests a detailed scheme of steps that can be followed during the test development process of a high-stakes test. The important components given in the scheme can be seen in Figure 5.

A. Needs analysis
1. define the purpose of the test
2. conduct a preliminary investigation
3. collect primary data
4. collect secondary data
5. analyze target language use task and language characteristics
B. How to turn target language use tasks into test tasks
1. developing a test task
2. developing a blueprint for the test
3. options for test tasks
C. Test administration
1. test environment
2. personnel
3. procedures
4. scoring

Figure 5. A scheme of test development (Adapted from Douglas, 2010)

Douglas (2010) states that the essential aspect of test development is creating “technically sound and practically useful tests that are at the same time stress free, fair and relevant for the test takers”, and offering them the best possible conditions on which they can demonstrate their language ability (p.63). To this end, following the above steps proposed by Douglas (2010) would enable test developers to create such tests and testing situations.

In addition to these frameworks mentioned above for test development, in the previous sections of this chapter, two main frameworks, Weir’s (2005) socio-cognitive framework for validating listening tests and Field’s (2013) cognitive

processing framework of listening, have been presented as the basis for the validity claims about the test in this study. Weir's framework suggests that test development should actually follow steps of a validation study. That is to say, theory-based validity and context validity should provide the basis for the test development process, and thus for test specifications. After this step, the scores obtained from the test should be analyzed and statistical analyses of these scores should be carried out for the investigation of scoring validity. The next step is to collect evidence for criterion-related validity for the validity of our inferences and eventually for consequential validity to see the effects of the test on the individuals. As a whole, it can be argued that this framework guides test developers in an organized and comprehensive way. The other framework adopted in this study, Field's cognitive framework for listening also guides the test developers in terms of the theory-based and context validity claims, and thus the components of the test specifications. The operationalization of the listening construct, the practical steps taken and the essential decisions made about test development in the present study are discussed below according to Weir's validation framework and Field's listening framework.

Based on Weir (2005), the first step to take in test development is the analyses of construct frameworks explaining the construct to be measured. The theoretical frameworks explaining the construct to be measured should be taken into consideration to determine the sub-skills to be tested, the components of the test specifications, the scoring criteria and other important concerns regarding test administration. Field's (2013) framework for listening demonstrates a theoretical background for the listening construct as explained already in this chapter. In addition to the theoretical frameworks, the CEFR also provides test developers with specifications of each language skill so that test developers have a better

understanding of the sub-skills that learners can achieve across different proficiency levels. These specifications included in the CEFR enable test developers to align their tests with the predicted language proficiency levels and standardize their tests in this respect. Therefore, employing the CEFR while designing language tests guides test developers in a standard and practical manner. In addition to the analyses of listening frameworks and the CEFR, analyzing the teaching and testing materials of the target construct also helps test developers to see the practices in teaching and assessing the target construct and encourages them to consider the relevant aspects of their own tests.

After analyses of frameworks and materials, one of the most essential steps of test development is to write test specifications. Bachman and Palmer (2010) state that test specifications aim to demonstrate the overall structure of the assessment, with many details such as the description of the assessment, the specifications for each task, the procedures for determining a benchmark score, the way assessment records will be collected and reported, and the way the assessment will be administered. Test specifications, according to Bachman and Palmer (2010), provide guidelines for quality control and justifications for the tasks chosen and enable test developers to compare these tasks with the TLU tasks. Test specifications also help us produce comparable forms of the same assessment since they will be similar in terms of their content, structure, and the task by following the same blueprint (Bachman and Palmer, 2010). McNamara (2000) mentions that test specifications include such details as “the length and structure of each part of the test, the type of materials used in the test, the source of these materials, their authenticity, the response format, the test rubric, and how the responses are to be scored” (p.31). McNamara (2000) also maintains that test specifications are made up of the

instructions for creating the test; therefore, even if somebody else other than the test developer wishes to create other versions of the test, they can do so by following the instructions in the test specifications. As a result, test specifications that we are aiming at producing should be prepared in the light of cognitive and validity frameworks so that there will be a systematic correspondence with the construct and its operationalization in the test tasks.

After the preparation of the test specifications, the test can be developed according to the criteria in the specifications and items can be written. One essential issue in test development is the task type and the response method required by the task. Buck (2001) offers a discussion on the three important approaches to assessing listening - discrete-point, integrative and communicative approaches- and examines the types of tasks that can be employed for each approach. Moreover, he gives examples of specific listening tasks and demonstrates key issues for developing items for listening tests. Similarly, Elliott and Wilson (2013) review the task types used in testing listening and discuss their pros and cons in terms of measuring the listening ability. After the evaluation of the positive and negative features of the task types, the items can be written by considering with the cognitive and contextual requirements of the construct and the proficiency levels.

Another crucial aspect of test development is the linguistic features of the tasks. While writing the items and the listening scripts, test developers should pay attention to the target proficiency level and the difficulty of the texts. Buck (2001) includes the text characteristics that impact on difficulty together with linguistic characteristics among explicitness, organization, content and context. Weir (2005) also lists a variety of aspects that impact the linguistic difficulty of tests, which are discussed thoroughly in section 2.3.3.3.

Another facet of test development is getting expert opinion (Fulcher & Davidson, 2007). Expert judges should be asked to evaluate the test in terms of whether it contains a sufficiently representative sample of the tasks from the TLU domain to be tested. These judges may be language teachers who have experience in teaching the target language or researchers who have expertise in teaching the language, or developing language tests. Their feedback should be integrated into the test development process to succeed in creating language tests as representative of the TLU domain as possible.

After receiving feedback and making any modifications if necessary, the test can be administered. Fulcher and Davidson (2007) state that the constraints that are likely to affect a successful test administration should be determined in advance and taken into consideration in related documents. These constraints can occur concerning people, skills, equipment, accommodation, security, information technology and money (pp.128-9). Furthermore, the parts of test specifications with regard to presentation of the test, scoring and interlocutors should be considered as well in order to administer a test successfully and in a standardized fashion.

After all the necessary steps are followed, another group of validity evidence comes from scoring data. According to Weir's framework for language validation, after the test has been delivered, the test scores should be analyzed to provide evidence for our scoring validity claims. Depending on the results of the statistical analyses, modifications in the tasks can be made. Buck (2001) states that test development is a cyclical process and after writing specifications, and developing and trying out tasks, test developers gain more information about how the test tasks work and how well they measure the construct. Therefore, test writers can review and revise test specifications, tasks, items and listening texts until they develop a test

that assesses the construct most appropriately and effectively. All this evidence put together support adequacy and accuracy of our interpretation and use of test scores.

2.5 Conclusion

In this chapter, we reviewed essential topics in test development; validity, a test validation framework, a listening framework and the steps in test development. In terms of test validation, the socio-cognitive framework for validating language tests by Weir (2005) was explained in detail, as it forms the theoretical rationale behind test development and validation in this study. Moreover, Field's (2013) cognitive framework for listening was explained comprehensively since it reflects the theoretical basis for the listening construct in the current study. Furthermore, some basic steps to be followed during test development based on the components of Weir's framework were discussed. Following the suggestions discussed above, the present study aims at providing validity evidence through the investigation of the following research questions:

1. What are the cognitive requirements of the listening test tasks?
 - a. Is the listening construct operationalized in the test tasks in a way that targets a sufficient range of cognitive processes indicated by the listening frameworks across different proficiency levels as predicted by the CEFR?
 - b. Do the test takers' perceptions of the listening sub-skills that they employed to answer the items support that the test tasks can tap into the necessary cognitive processes?
2. What are the contextual characteristics of the listening test tasks?

- a. What are the demands imposed upon the test takers by task setting, administration setting, linguistic features of the listening test tasks and the speakers?
 - b. What are the participants' perceptions of the tasks in terms of the suitability of their contextual features for the different proficiency levels?
3. How well do the test tasks and the items function in terms of scoring validity?
- a. Do the values for central tendency measures of the tasks and item analyses based on the test takers' performances support that the test is functioning well?
 - b. Does the test measure the listening ability of learners of TFL reliably?

The methods adopted to investigate these research questions are explained in Chapter 3 in detail and the argumentation and empirical findings on these research questions are discussed in Chapter 4.

CHAPTER 3

METHODOLOGY

3.1 Introduction

The purpose of this chapter is to describe the methods employed for the investigation of the research questions in this study. In this chapter, details concerning the participants, the instruments, the test development procedure and the research questions and their data analysis are presented. It should be noted again that this study aims at developing a listening test for learners of TFL and providing evidence for the validity claims made based on the test scores of the test takers. The framework adopted for test validation and test development in this study is the socio-cognitive framework for validating listening tests proposed by Weir (2005), as discussed in Chapter 2 in detail. The components of this framework to be investigated in this study are theory-based validity, context validity and scoring validity. The research questions in this study seek answers for the justification of the claims made on these validity types.

3.2 Participants

3.2.1 Participants in the first administration

The participants of the first piloting were a group of 55 students who came to Boğaziçi University via the Erasmus Exchange Program in the fall semester of 2014. These students took Turkish for Foreigners (TKF) classes offered by the Department of Turkish Language and Literature at Boğaziçi University. The classes that they were taking were named as TKF 211, TKF 315 and TKF 317. In TKF 211 course,

the proficiency level of the students were assumed to be at intermediate level while the students in TKF 315 and 317 classes were considered to be at upper-intermediate level. According to the feedback received from the instructors who deliver these courses, the classification of the students to the levels is not usually carried out via a standardized test; rather, intuition of the instructors (formed through interviews or e-mail exchanges with the students) and the students' own perceptions of their language proficiency are the factors that determine students' classes.

3.2.2 Participants in the second administration

The second version of the test was administered to a group of 30 learners of TFL at Boğaziçi University. The participants attended Turkish Language and Culture Program (TLCP) Summer Course in the summer of 2016 offered by the Language Center at Boğaziçi University. The courses offered during the summer course are named as 20, 21, 25, 30 and 31. These classes are not CEFR aligned; however, according to the instructors, their presumed CEFR alignments are A1+ and A2 for Class 20, A2 and B1 for Classes 21 and 25, and B2 and B2+ for Classes 30 and 31. The instructors maintained that they utilized a test that they developed to place the students into these classes.

In the second piloting, the test takers were divided into two groups as lower-level test takers and higher-level test takers. Not all the test tasks were administered to the entire group of test takers due to the restrictions brought by the class instructors. The lower-level test takers were in Classes 20, 21 and 25 and took the test tasks at A1, A2 and B1 levels. The higher-level test takers were in Classes 30 and 31 and took the test tasks at A2, B1 and B2 levels. Therefore, the data collected

in the second administration were analyzed separately for these two groups of test takers and the results were reported individually as well.

3.3 Instruments

3.3.1 Tasks

In the first administration of the test, the data were collected through the administration of the five listening test tasks prepared by the researcher. These tasks ranged from A1 to C1 CEFR levels and included a variety of question types such as gap-filling, multiple-choice and short-answer questions (See Appendix B for the tasks in the first piloting). As a result of the modifications made after the first piloting, only four test tasks at A1, A2, B1 and B2 levels were delivered in the second administration of the test because the task at C1 level had been removed from the second version of the test for technical reasons that are explained in section 4.2.5 in Chapter 4 (See Appendix C for the tasks in the second piloting). In the current study, the tasks provided substantial data regarding the items, the listening texts, the recordings of the listening texts and the test scores obtained from the test takers.

3.3.2 Test specifications

The test specifications prepared for the development of the test tasks were also utilized to gather validity evidence (See Appendix D for the test specifications). The specifications were prepared for each listening test task separately. They include task, text and item characteristics detailed in terms of the cognitive, contextual and scoring requirements of each test task. The features of the test tasks outlined in the

test specifications were examined in terms of their suitability for theoretical frameworks as discussed in detail in sections 3.5.1 and 3.5.2 below.

3.3.3 Task evaluation questionnaires

The task evaluation questionnaires were given to the test takers in the second administration of the test (See Appendix E for the task evaluation questionnaires). The test takers were asked to complete the task evaluation questionnaires designed specifically for each task. As soon as the test takers finished answering the questions for a specific task, they immediately completed the questionnaire for that task.

The task evaluation questionnaires are composed of three sections. In the first section of the questionnaires, the test takers were required to evaluate the listening sub-skills they employed to answer each item in the test tasks. In this section of the questionnaires, the test takers were given a list of listening sub-skills compiled from the theoretical frameworks by Field (2013), Weir (1993) and Richards (1983). The test takers were asked to choose and mark the listening sub-skills; i.e. the cognitive processes they used while answering the items in the tasks. They could choose as many sub-skills as they wanted for each of the test items. The findings for the first section were reported by showing the number of test takers who chose each sub-skill in tables. The most popular sub-skills chosen by the test takers were also marked with an asterisk (*) for emphasis. In the second section of the task evaluation questionnaires, the test takers evaluated the difficulty level of each item in each test task on a scale ranging from 1 (*too easy*) to 4 (*too difficult*). The mean scores of the test takers' answers are calculated to determine the perceived difficulty levels of the items. The third section of the task evaluation questionnaires informs the researcher about the test takers' perceptions on the contextual appropriacy of the tasks for a

broader validity argument. The test takers were asked to complete the likert-scales for a number of statements about the contextual characteristics of the tasks and the items.

3.3.4 CEFR specifications

The CEFR specifications (the Council of Europe, 2001) offer descriptions of the listening abilities of language learners at various proficiency levels (See Appendix A for the CEFR specifications for the overall listening ability). These specifications were used as a point of reference for the development of the test specifications in this study in terms of cognitive requirements. In addition, the Reference Level Descriptors (Breakthrough (Trim, 2009), Waystage 1990 and Threshold 1990 (van Ek and Trim, 1991), and Vantage (van Ek and Trim, 2001) prepared at A1, A2, B1 and B2 levels respectively based on the CEFR provided the basis for the contextual features of the test tasks in the current study.

3.4 Procedure

For the development of the current test, firstly, a test validation framework, Weir's (2005) socio-cognitive framework for test validation was adopted in order to determine the steps to be followed during test development. Furthermore, Field's (2013) listening framework and the CEFR descriptors for listening ability were employed in order to define the listening construct. In addition to the theoretical frameworks, course books and syllabi of Turkish for Foreigners courses offered at Boğaziçi University, Hitit course books for foreign language learners of Turkish, and syllabi of Turkish for foreigners courses at universities in Turkey and abroad were examined in order to attain a better understanding of the grammatical structures,

topics and vocabulary that are studied at different levels of proficiency for TFL. After the analysis of the theoretical frameworks and course materials, the test specifications were created and the important parameters in the theoretical frameworks were taken into consideration during the development of the test specifications. After this step, the tasks were developed by the researcher based on the criteria in the test specifications. The first version of the test was administered in fall semester of 2014 at Boğaziçi University to a group of 55 participants as explained in section 3.2.1. The test was administered by the instructors in the classroom environment. Before taking part in the study, the participants signed a consent form showing their willingness to participate in the study (See Appendix F for the consent form). During the administration, the participants were asked to complete the tasks and the task evaluation questionnaires in the allocated time period, which mostly lasted one class hour (45-50 minutes). After the first administration, statistical analyses were conducted as discussed in section 3.6.3. As a result of the statistical analyses conducted after the first piloting, some problematic items and the tasks were revised or rejected to develop the second version of the test. Although it was observed that the statistical values within the tasks themselves were generally good after the first piloting, some changes were still necessary to increase cognitive and contextual relevance. Certain modifications had to be made to adjust the difficulty levels of the tasks as well. Therefore, test revisions based on statistical findings, in addition to the modifications necessitated by construct and context-related issues, had to be made with a testing expert as discussed thoroughly in section 4.2. After the necessary revisions were made by the researcher and the testing expert, expert opinion was also gathered from the TFL instructors at Boğaziçi University. A few further modifications were also made based on the suggestions received from the

experts. The second version of the test was administered in summer school of 2016 to 30 participants as mentioned in section 3.2.2. The test tasks were given to the participants in their class hours by their instructors in the allocated time period (45-50 minutes) and consent forms were also collected from the participants. After the second administration, the test scores of the test takers were analyzed again following classical item analysis procedures. While analyzing the data for the second piloting, separate analyses were conducted to show the results obtained from the different groups of test takers. This can be seen as a limitation of this study; however, removing participants from the study was not an option and therefore, this procedure had to be followed.

3.5 Data analysis

3.5.1 Research question 1

The first research question to be examined in the study is stated below:

1. What are the cognitive requirements of the listening test tasks?
 - a. Is the listening construct operationalized in the test tasks in a way that targets a sufficient range of cognitive processes indicated by the listening frameworks across different proficiency levels as predicted by the CEFR?
 - b. Do the test takers' perceptions of the listening sub-skills that they employed to answer the items support that the test tasks can tap into the necessary cognitive processes?

This research question aims to find evidence for the theory-based validity claims of the listening test in this study in terms of construct relevance and representativeness.

In other words, it investigates the similarity of processing and comprehensiveness in the test tasks (Field, 2013, p.80).

In order to investigate the theory-based validity claims of this test, a variety of instruments were employed. Firstly, the listening sub-skills specified in the test specifications for both administrations are examined in terms of their alignment with the listening framework proposed by Field (2013), which forms the major theoretical basis for the listening construct in this test. Whether the target sub-skills in the test specifications for each task match the sub-skills in this listening framework is investigated. Furthermore, the CEFR specifications for the listening skill across different proficiency levels, and Weir's (1993) and Richard's (1983) listening frameworks, which specify essential listening sub-skills, are also taken into consideration for the investigation of the alignment of the target skills in the test tasks with the sub-skills included in these frameworks. Thus, a comprehensive reference is made to the theory on the listening construct. The analysis of the listening sub-skills in the test specifications according to theoretical frameworks is considered to demonstrate whether the tasks tap into a sufficiently broad range of cognitive processes. The diversity of the cognitive processes targeted by the test tasks will also enable us to see the extent to which construct representation is achieved in the study.

Secondly, the test takers' responses to the first section of the task evaluation questionnaires are reported as stated in section 3.3. The most popular cognitive processes chosen by a majority of the test takers are examined in terms of their congruence with the target sub-skills of the tasks as determined in the test specifications and the sub-skills specified in the listening frameworks mentioned above. This procedure aims to provide retrospective evidence for the relevance of the

cognitive processes purportedly elicited by the test tasks from the perspective of the test takers.

3.5.2 Research question 2

The second research question and its sub-questions that investigate issues and validity claims related to context validity are stated below:

2. What are the contextual characteristics of the listening test tasks?
 - a. What are the demands imposed upon the test takers by task setting, administration setting, linguistic features of the listening test tasks and the speakers?
 - b. What are the participants' perceptions of the tasks in terms of the suitability of their contextual features for the different proficiency levels?

According to Weir (2005), essential components of cognitive validity are divided into three main categories as task setting, administration setting and task demands (linguistic and interlocutor). He also mentions several sub-topics for each of these categories and states that each component of context validity needs to be given careful thought during test development as discussed in section 2.3.3. For the discussion of the essential components that should warrant context validity, Elliott and Wilson's (2013) study on the application of Weir's (2005) framework to Cambridge tests is followed in this study, as it offers a comprehensive and organized investigation of this validity type based on Weir (2005). Elliott and Wilson (2013) investigate the context validity of the Cambridge ESOL suite's listening tests and provide evidence for the validity claims based on the scores from these tests. Following Weir (2005) and Elliott and Wilson's (2013) discussion on the socio-cognitive framework for validating listening tests, the contextual characteristics of

the current test are investigated in Chapter 4. The results of this investigation are given in the form of a discussion on the appropriacy of the contextual characteristics implemented in the present test according to the theoretical frameworks mentioned above.

For the investigation of the second research question, firstly each component of context validity is discussed with reference to the test specifications and also the tasks in the first and second administrations. The test specifications include details regarding contextual characteristics of the tasks; therefore, a detailed justification of the selection and control of the contextual features as specified in test specifications is necessary. Issues related to task setting are explored via the justification of the choices and decisions made during test development by following Weir (2005) and Elliott and Wilson (2013). Text purpose, response format, known criteria, weighting, order of items and time constraints are discussed thoroughly with reference to the test specifications and the tasks (items, listening texts and recordings). In addition to the task setting, administration setting of the current task is also explored in terms of physical conditions, uniformity of administration and security. The third aspect of context validity, the demands of the tasks are scrutinized in two categories; linguistic and interlocutor demands. The linguistics demands of the tasks and the items are explored according to the specifications of the CEFR levels as given in Reference Level Descriptors (Breakthrough (Trim, 2009), Waystage 1990 and Threshold 1990 (van Ek and Trim, 1991), and Vantage (van Ek and Trim, 2001)). Under the investigation of linguistic demands, the functional languages, the grammatical structures and the vocabulary used in the items and the listening texts are investigated in terms of their alignment with the specifications in the Reference Level Descriptors. Specific examples of functional languages, grammatical structures

and vocabulary from the listening texts and the items are given as well to demonstrate evidence for this alignment. For the other components of the linguistics demands such as discourse mode, channel, text length, nature of information, and content knowledge, the relevant parts of the test specifications and the listening texts are addressed and their appropriateness in terms of context validity is also discussed. Finally, issues regarding interlocutors such as speech rate, variety of accent, acquaintanceship, and the number and gender of speakers are investigated through the discussion of the related parts of the test specifications, the listening texts and the recordings. Speech rate is also discussed via word per minute (wpm) calculations for each recorded text. The calculations are evaluated in terms of their appropriacy for context validity and different proficiency levels.

In addition to the theoretical analysis of the current test based on Weir (2005) and Elliott and Wilson (2013), the data obtained from the third section of the task evaluation questionnaires are also utilized to show the test takers' perceptions on the various components of context validity. In the third section of the task evaluation questionnaires, the test takers evaluated a number of statements related to different aspects of context validity such as the clarity of the rubrics, the audibility, comprehensibility and speed of the recordings, the times of playing the recording, the authenticity and the difficulty level of the recorded text, and the difficulty level of the tasks. The mean scores for each of these statements are calculated and the test takers' perceptions of the appropriacy of the contextual features of the tasks are demonstrated through these mean scores. The collection of this data enabled the researcher to support the theoretical discussion with statistical evidence collected from the test takers.

3.5.3 Research question 3

The research question stated below has been formulated to investigate scoring validity.

3. How well do the test tasks and the items function in terms of scoring validity?
 - a. Do the values for central tendency measures of the tasks and item analyses based on the test takers' performances support that the test is functioning well?
 - b. Does the test measure the listening ability of learners of TFL reliably?

Weir (2005) states that collecting evidence for scoring validity provides test developers with a posteriori evidence. The investigation of this validity type demonstrates how effectively the test tasks and the items work in terms of central tendency, reliability and item characteristics.

The investigation of the third research question is conducted through the analyses of the test scores obtained from the first and second piloting studies and the second section of the task evaluation questionnaires. Firstly, the items were marked by the researcher dichotomously, by giving "1" for the correct answers and "0" for the incorrect answers. After marking, the scores are analyzed on IBM SPSS 21 Software by employing classical item analysis procedures (measures of central tendency, item analysis and reliability). Central tendency measures such as mean, median, standard deviation, skewness and kurtosis assist test developers to have a summarizing and meaningful description of the test scores obtained from tests or other measures. They demonstrate an overall picture of how the test scores are distributed and grouped (Bachman, 2004). In this study, mean, range, and standard deviation values of the test scores are calculated in order to demonstrate measures of central tendency and dispersion of the test scores across tasks. Moreover, values for skewness and kurtosis- peakedness- are also calculated. These values can be positive,

negative or centered around zero. For skewness, a value of zero means that the scores are distributed symmetrically while negative and positive scores show negative and positive skewness, respectively. In terms of kurtosis statistics, a value of zero indicates normal distribution, negative statistics indicate flat distributions and positive statistics indicate peaked distributions. It is generally accepted that the values for skewness and kurtosis between -2 and +2 demonstrate a reasonably normal distribution (Bachman, 2004). All of these values together provide a more comprehensive picture of the way the scores are distributed on a continuum.

As to the reliability of criterion-referenced tests such as the one under investigation, Bachman (2004) states that index of dependability should be calculated to determine the extent to which the test scores are reliable as an indicator of the test taker's mastery level of the target skill. The type of statistical data to be collected with respect to reliability in this study is the internal consistency coefficients due to the test design and practicality. One form of internal consistency coefficient is Kuder-Richardson 20, which can be used to analyze reliability for dichotomously scored items (scored either 1 or 0) (Bachman, 2004). However, Cronbach's alpha (α) is a more general estimator of reliability (Bachman, 2004) and it is more commonly quoted when reporting estimates of reliability (Geranpayeh, 2013). Therefore, values for Cronbach's alpha (α) are demonstrated to estimate and report reliability in this study.

Item characteristics, on the other hand, provide information with respect to the discrimination, difficulty and reliability values of the individual items and enable test takers to detect the problematic items and areas in a test by showing a statistical value for each. For item analysis, values for item facility (IF), item discrimination (ID; Corrected Item-Total Correlation-CITC), and reliability estimates for individual

test items (Alpha If Item Deleted; AIID) are demonstrated to evaluate the effectiveness of the individual items. IF shows the proportion of the people who answered an item correctly (Fulcher & Davidson, 2007). It is suggested that the items should not be too easy or too difficult for the target group of test takers. Therefore, an IF value of 0.5 is considered to be reasonable and the values between 0.3 and 0.7 are deemed acceptable as well (Henning, 1987, in Fulcher & Davidson, 2007). The lower the IF value is, the more difficult the item is for the test takers. Another component of item analysis, ID indicates the ability of the individual items to discriminate between higher and lower-level test takers. The acceptable values for ID should be higher than .30, and the items with an ID value below this should either be revised or omitted from the test (Fulcher & Davidson, 2007). In addition to all these analyses, the internal consistency measures for the individual items should also be calculated to see how much each item contributes to the internal consistency of the test. AIID tells us if the alpha score for the test would increase or decrease if a specific item is omitted from the test. If the general alpha increases when the item is deleted, it means that the item does not contribute positively to the reliability of the test and therefore, should be either modified or removed from the test.

In addition to the analysis of the test scores, the second section of the task evaluation questionnaires given in the second administration also provided valuable information with respect to the perceived difficulty levels of the items and tasks as mentioned in section 3.3. The findings from the second section of the task evaluation questionnaires are considered to help test developers make the necessary changes in the items and the tasks and create the final version of the test. Furthermore, it enables the test developers to see whether the data collected from the test takers support the statistical findings.

3.6 Conclusion

This chapter aimed to explain the methods to be used in the exploration of the research questions in the current study. These research questions attempt to gather validity evidence in terms of cognitive, context and scoring validity components of Weir's (2005) socio-cognitive framework for validating language tests. As mentioned earlier, the other important components of the framework, test taker characteristics, criterion-related validity and consequential validity are beyond the scope of this research and can be investigated through further research studies in the future after the genuine administration of the test for measuring proficiency levels of the future students of TFL.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

As mentioned in the previous chapters, this study is conducted to develop an academic listening test for learners of TFL and collect a priori and a posteriori evidence for our validity claims based on the characteristics of the test tasks and the scores of the test takers. Therefore, following Weir's (2005) socio-cognitive framework for validating listening tests and Field's (2013) cognitive framework of listening, theory-based validity, context validity and scoring validity claims made about this test are discussed in this chapter, and validity evidence is gathered to justify our claims based on these frameworks. To this end, three research questions have been formulated and they are discussed respectively here as specified in Chapter 3. Moreover, discussions and suggestions are made regarding the future versions of the test based on the findings from the analysis of the data and the discussions in this chapter.

4.2 Investigation of theory-based validity

The first research question seeks evidence to support the inferences made about theory-based validity. The question is formulated as follows:

Research Question 1: What are the cognitive requirements of the listening test tasks?

- a. Is the listening construct operationalized in the test tasks in a way that targets a sufficient range of cognitive processes indicated by the listening frameworks across different proficiency levels as predicted by the CEFR?

- b. Do the test takers' perceptions of the listening sub-skills that they employed to answer the items support that the test tasks can tap into the necessary cognitive processes?

For the investigation of the first research question, the test specifications for each test task are scrutinized according to Field (2013), the CEFR descriptors for the listening skill, Weir (1993) and Richards (1983). Furthermore, the results of the first section of the task evaluation questionnaires are discussed below for each task in the second administration.

4.2.1 A1 level task

A1 level tasks given in both the first and the second administration are discussed below in terms of the cognitive processes they targeted.

4.2.1.1 Cognitive requirements of A1 level task according to theoretical frameworks

A1 level task in the study aimed to assess one listening sub-skill, which was “to listen for specific factual information clearly stated”. In A1 level task, the test takers were supposed to answer six open-ended questions with short answers such as numbers or one or two words. This task aims at processing specific word-level information, thus lexical search as the cognitive process. The answers are also clearly stated and pronounced in the listening recordings. Therefore, the test takers are assumed to be able to respond to the questions using lexical information only. Based on this, it can be stated that A1 level task, which is designed as the easiest task in the test, measures one of the lower-level processes in Field's framework and therefore, complies with the cognitive requirements of the listening framework at this level. It is worth noting at this point that none of the tasks specifically targeted input

decoding, the lowest level of processing in Field's framework. The reason for this is that this process is thought to be achieved at all proficiency levels and assessing the distinction between sounds via minimal pairs is a question type that was asked in early stages of testing listening (Brown, 1990). The communicative view of language learning and testing does not support this question type very much (Taylor, 2013). It is also assumed that even in the lowest level task in the test, this process is tapped, as it is a prerequisite for even the lowest level tasks. Therefore, it is not included in the test as a target sub-skill.

With respect to the CEFR specifications at A1 level, the global scale is that the listener "can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type." In addition, for overall listening comprehension at A1 level, it is stated that the listener "can follow speech that is very slow and carefully articulated, with long pauses for him/her to assimilate meaning" (See Appendix A for CEFR specifications for overall listening comprehension). These specifications indicate that at A1 level listeners are assumed to understand simple, high frequency vocabulary when uttered slowly and clearly. The answers for the A1 level task in this study are composed of simple and high frequency words, too and the listening text was recorded at a slow and understandable speed. In addition, in the CEFR scales for writing, it is stated that learners at A1 level "can write simple isolated phrases or sentences". Therefore, short-answer questions are considered to be an appropriate means of assessing this skill.

Similarly, Weir (1993) mentions the sub-skills for direct meaning comprehension in his taxonomy and this list of sub-skills also includes listening for specific information. Based on the findings from Field (2013), the CEFR

specifications and Weir (1993), it can be concluded that this task is designed to measure only lower-level processes and understanding, factual, simple and clear information. In addition, in terms of construct representation, targeting only this sub-skill can be seen sufficient at the lowest level of proficiency.

However, although the task is considered to assess the target sub-skill, one aspect of the test task that underwent changes was the nature of the information the items required. In the first examination, there were four questions which needed numerical answers and two questions which demanded content words (See Appendix B for the first A1 level task). In order to create a balance, one of the questions was replaced by a question that required comprehension of a content word (See Appendix C for the second A1 level task). By doing this, the test writers (the researcher and the testing expert) aimed to achieve a better construct representation, by still targeting comprehension of clear lexical information.

4.2.1.2 Cognitive requirements of A1 level task according to task evaluation questionnaires

The first section of the questionnaire for A1 level task was composed of nine sub-skills, from which the test takers selected for the six items in the task (See Appendix E for the task evaluation questionnaires). For this task, the results of the questionnaires were calculated for the lower-level students (n=16), as only this group took this task. The total numbers of the test takers who marked each sub-skill for each item were given in Table 1. The sub-skills preferred by the test takers most are shown with an asterisk (*) next to the numbers.

Table 1. Lower-level Test Takers' Perceptions of Cognitive Processes in A1 Level Task (n=16)

In order to answer this question correctly I had to...	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1. understand specific bits of information in the dialogue	11*	13*	13*	13*	9*	9*
2. understand just the main idea(s)	6*	3	3	3	7*	6*
3. understand the details used to explain the main idea(s)	4	3	3	3	5	4
4. differentiate between important and less important information	11*	10*	6*	6*	7*	6*
5. understand what the dialogue is about briefly	9*	6*	6*	5	8*	6*
6. understand how information in the whole dialogue fits together	3	4	3	3	7*	4
7. pay attention to the speakers' attitude and tone	2	2	3	2	4	3
8. understand what the speaker's intention is when using a certain sentence	3	3	1	1	2	3
9. rely on my general world knowledge	3	1	0	0	4	6*

Table 1 shows that the most popular sub-skill marked by the test takers for all of the items is the first sub-skill “understand specific bits of information in the dialogue”, which is what this task precisely aims to tap into. In addition to this, there are some important implications in this table regarding the nature of the items. For the first item, the first and the fourth sub-skills were marked by 11 test takers. This shows that most of the test takers needed to employ an additional cognitive process while responding to this item. The reason for the utilization of another sub-skill may stem from the nature of this question. In the first item, the test takers were supposed to complete a list of things which were bought at a stationery shop and therefore, they may have focused only on the items related to this list by selecting the information available in the text. They also marked the fifth sub-skill, which demonstrates that the test takers attempted to understand the overall idea of the text.

Similarly, the numbers of test takers who marked the first and the fourth sub-skills for the second item are very close to each other. For the second item, test takers

needed to answer an open-ended question by writing the total price for the goods bought in the stationery shop. While answering this question, the test takers may have only listened for information that is related to numbers and prices by ignoring the other irrelevant information in the listening text. As a result, although the fourth sub-skill in the questionnaire is not specifically targeted by the first and second items, it seems logical to employ this sub-skill while answering these questions. It can be an artifact of test taking condition that the test takers had to search for the right answer. This can also be considered as a strategy adopted by the test takers, although the item can be answered without employing this sub-skill. Therefore, adoption of this sub-skill is not deemed as a major threat to the cognitive validity claims about the test and since the focus of the first and second items is to understand lexical information, they can be considered successful in measuring this sub-skill.

For the third and fourth items, there are no prevailing sub-skills chosen by the test takers other than the target sub-skill. This indicates that these items succeed in eliciting the target sub-skill. For the fifth item, the visible sub-skills other than the target one are generally related to the organization of ideas such as the second, fourth, fifth and sixth sub-skills. There seems to be an equal distribution of the chosen sub-skills among the second, fourth, fifth and sixth sub-skills. For the sixth item, similarly, the second, fourth and fifth sub-skills received a close number of markings to that of the target sub-skill. In addition, for this item, the ninth sub-skill also was preferred by a high number of test takers, too. The situation for the fifth and sixth items may indicate that these items may have posed different cognitive loads on the test takers. These items are related to a short dialogue between a doctor and a patient, and the test takers respond to such questions as “What part of the patient’s body hurts?” and “What did the doctor give to the patient?”. Therefore, they could

have relied more on the organization of the text to find the related areas to the questions. Moreover, the sentences become relatively longer in this short dialogue and the content load is heavier when compared to the other dialogues in the task. Thus, the test takers may have needed to utilize more listening sub-skills to understand the text. Another important implication here is the use of general world knowledge (the ninth sub-skill), especially for the sixth item. The item asks what the doctor gave to the patient and the test takers might have relied more on their world knowledge to predict and listen for the correct information, which is “(ağrı kesici) ilaç” (pain killer / medicine). To prevent such a case, the answer can be revised as one that cannot be predicted with common sense.

In short, it can be concluded about this task that the target sub-skill, which corresponds to the first sub-skill in the questionnaire, is elicited from the test takers successfully although some extra sub-skills such as the fourth and fifth sub-skills are also employed. The listeners might have felt the need to employ additional cognitive processes in order to reach an effective understanding of the listening text and to compensate for their lack of linguistic knowledge. Therefore, as long as the items can successfully elicit the target sub-skills, using other sub-skills should not be considered as a threat to the cognitive validity of the claims we make about this test and the test results.

4.2.2 A2 level task

A2 level tasks given in both the first and the second administration are discussed below in terms of the cognitive processes they targeted. The results of the task evaluation questionnaires are also presented individually for two different groups of test takers.

4.2.2.1 Cognitive requirements of A2 level task according to theoretical frameworks

The task at A2 level was modified dramatically after the first administration. This task was originally designed as a B1 level task; however, after analyzing the characteristics of the test task after the first piloting, it was changed into an A2 level task (See Appendix B for the first B1 level task and Appendix C for the second A2 level task). One reason for this radical change was that the cognitive processes elicited by the items did not meet the cognitive requirements of the target proficiency level and the target processes in the frameworks. Another reason was the results of the item analysis and statistics; however, they are not discussed here, but in the investigation of research question 3 in section 4.4 in this chapter.

On the other hand, the first A2 level task was also problematic due to cognitive concerns and statistical analysis results and thus, it was later utilized as B1 level task. Due to these changes, we have to make references to both tasks to compare their previous and final versions, and discuss the reasons for the alterations made on the first versions. In the following pages, “A2 level task in the first administration” is also referred to as “B1 level task in the second administration”. Likewise, “B1 level task in the first administration” is referred to as “A2 level task in the second administration”.

The target skill aimed to be measured in A2 level task in the first administration was to “listen for specific information”. In this task, the test takers were supposed to listen to a conversation between two classmates and answer some multiple-choice questions. The questions were assumed to require comprehension of specific details in the listening, and thus, lexical and sentence-level factual information. These two types of information can be related to the lower-level processes in Field’s (2013) framework; lexical search and parsing. At this

proficiency level, expecting lower-level processes from the test takers seems logical and therefore, it can be stated that the task seemed in line with the framework in this respect.

As to the CEFR scales for overall listening comprehension at A2 level, learners are expected to “understand phrases and expressions related to areas of most immediate priority (e.g. very basic personal and family information, shopping, local geography, employment) provided speech is clearly and slowly articulated” and “understand enough to be able to meet needs of a concrete type provided speech is clearly and slowly articulated”. These statements mean that at A2 level learners are expected to understand direct and clear messages about everyday topics that they are familiar with. The task is considered to meet the requirements of the CEFR specifications in this aspect.

However, when the items were analyzed in terms of their cognitive demands after the first piloting, it was realized that some revisions were needed. Although most of the items seemed to assess the target skills, unfortunately, some of the items targeted cognitive processes inappropriate for this level. For certain items in A2 level task in the first administration, especially the first, third and sixth items, some of the answers were not directly stated in the listening text and the test takers were required to make inferences using the information in the text. Inferencing is not considered as a suitable target skill in the CEFR specifications for A2 level, as listeners are expected to understand “clearly stated information”. In addition to the CEFR scales, according to other listening frameworks, inferencing is considered to be a higher-level process happening during meaning construction (i.e. Field, 2013) and it comes after direct meaning comprehension in Weir’s (1993) list of listening sub-skills. These sources also provided evidence for the need for some modifications in this

task. Moreover, the statistical analyses conducted after the first piloting demonstrated that the task was above the expected level of difficulty for the test takers. Thus, this task was changed to be at B1 level in the second piloting after some modifications were made.

After the original A2 level task was modified, the new A2 level task in the second piloting also required some changes, as it was previously B1 level task. The cognitive skills targeted at this level were revised and in addition to “listening for specific factual information clearly stated”, another cognitive skill was added from Weir’s (1993) taxonomy, direct meaning comprehension section: “Listening for main idea(s) or important information: and distinguishing that from supporting detail, or examples”. The reason for this was to take the task beyond the A1 level task, and to test other sub-skills as well according to the requirements of the proficiency level. The original version of this task, B1 level task in the first administration, contained questions that demanded comprehension of lexical factual information. The test takers were asked to listen to a conversation between three friends about an introductory course and fill in some blanks with the information from the listening text. In the new version of the task, all of the questions were changed although some of the answers stayed the same. Instead of completing the notes, test takers were asked to write one-word answers for open-ended questions, choose the correct answer, and fill in the gaps in the sentences. The variety in the questions resulted from the aim to test different kinds of information and tap into different levels of processing. For instance, the multiple-choice question was created to assess sentence-level factual comprehension instead of only assessing lexical information. The other questions were written in short-answer format to make the type of information required clearer and suit more to the proficiency level. In addition, as

opposed to its previous version only one question that needed a numerical response was included in the new task in order to create a balance in terms of the kind of information required.

However, despite the revisions made for the target cognitive processes, when the types of information required by the items was analyzed again after the second administration, it was seen that almost all of the questions focus on lexical information again, as the items require only one-word answers and only the first item (the multiple-choice item) seems to require a cognitive process above lexical search. The first item does not only require sentence-level comprehension, but also inferencing, which is considered to be a higher-level process both in Field's (2013) and Weir's (1993) frameworks and is not stated in the CEFR scales as a sub-skill to be achieved at A2 level. Therefore, this question should be revised to make it more suitable for the cognitive requirements of the frameworks and transformed into an item that requires only sentence-level factual information. Furthermore, a few more items can also be modified to measure sentence-level factual information in this task so that A2 level task can be differentiated from A1 level task clearly without being beyond the assumed proficiency level.

4.2.2.2 Cognitive requirements of A2 level task according to task evaluation questionnaires

This task was administered to both lower-level and higher-level test takers in the second administration as explained in section 3.2.2. Therefore, the results of the task evaluation questionnaires for this task are presented in two separate tables for each group of test takers. In the task evaluation questionnaire, there were nine sub-skills to indicate the ones that they used while answering the eight items (See Appendix E for

the task evaluation questionnaires). The results for both groups of participants can be seen in Table 2 and Table 3. The most popular sub-skills preferred by the test takers are indicated with an asterisk (*).

Table 2. Lower-level Test Takers' Perceptions of Cognitive Processes in A2 Level Task (n=16)

In order to answer this question correctly I had to...	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
1. understand specific bits of information in the dialogue	11*	15*	16*	16*	15*	15*	14*	14*
2. understand just the main idea(s)	7*	2	2	1	1	2	2	1
3. understand the details used to explain the main idea(s)	4	4	4	4	3	5	7*	7*
4. differentiate between important and less important information	9*	7*	9*	8*	7*	6*	7*	6*
5. understand what the dialogue is about briefly	10*	5	6*	5	7*	4	4	3
6. understand how information in the whole dialogue fits together	4	0	0	2	1	1	3	1
7. pay attention to the speakers' attitude and tone	6*	2	2	2	1	1	2	1
8. understand what the speaker's intention is when using a certain sentence	3	2	2	2	1	1	2	1
9. rely on my general world knowledge	5	1	2	2	3	1	1	1

Table 2 shows that the lower-level group mainly employed the first and fourth sub-skills while answering the questions. The target sub-skill for this task was to listen for specific information at lexical level for items two to eight; thus, the items seemed to have elicited the necessary sub-skills since both the first and fourth sub-skills are related to comprehension of specific details. For the first item, the test takers also marked other sub-skills (the second, fifth and seventh sub-skills). This can show that the test takers might have had to use a combination of sub-skills. This item aimed to measure sentence-level factual information; however, it was realized after the second piloting that this item might have required some inferencing skills on the information in the text and on the speaker's attitude and tone. Besides, since this item is the first to appear in the test, the test takers may have attempted to form an

overall idea about the topic of the text and thus used the second and fifth sub-skills. For the seventh and eighth items, the lower-level test takers also tried to understand the details used to explain the main ideas (the third sub-skill). Utilizing this sub-skill is also sensible due to the nature of these items. There was one item stem for the seventh and eighth items and the test takers needed to listen to a relatively longer stretch of utterances in order to be able to understand the correct answers. The question asked “Dersin ödevleri nelerdir? Aşağıya yazınız.” “What are the requirements of the course? Write them below.” Therefore, they needed to follow the conversation and understand specific information, and thus used the third sub-skill.

According to Table 3, the higher-level test takers mostly employed sub-skills related to understanding specific information (the first and third sub-skills). As opposed to the lower-level group, the higher-level group did not depend much on differentiating between important and less important information, which was, indeed, not necessary to carry out the task, but instead, they focused on the details used to understand the main ideas. However, they similarly attempted to understand the topic of the text for the first item by employing the fifth sub-skill. These results demonstrate that the test takers generally adopted sub-skills which helped them understand specific information. This statistical finding supports the theoretical findings that this task heavily assesses specific lexical information. As mentioned previously, some of the items can be modified to target sentence-level factual information in the later versions of the test so that it can be differentiated from A1 level task. It is also important to underline that the findings related with this task can suggest that test takers from different levels can use different sub-skills to respond to a task.

Table 3. Higher-level Test Takers' Perceptions of Cognitive Processes in A2 Level Task (n=14)

In order to answer this question correctly I had to...	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
1. understand specific bits of information in the dialogue	7*	8*	12*	12*	11*	11*	10*	14*
2. understand just the main idea(s)	2	2	0	1	0	0	1	2
3. understand the details used to explain the main idea(s)	5*	5*	6*	5*	6*	6*	6*	6*
4. differentiate between important and less important information	4	3	5*	5*	3	4	3	3
5. understand what the dialogue is about briefly	7*	1	2	2	2	3	2	2
6. understand how information in the whole dialogue fits together	3	2	4	2	4	1	4	1
7. pay attention to the speakers' attitude and tone	0	0	0	0	1	1	0	0
8. understand what the speaker's intention is when using a certain sentence	1	0	0	1	1	0	0	0
9. rely on my general world knowledge	1	1	2	1	1	1	1	1

4.2.3 B1 level task

B1 level tasks given in both the first and the second administration are discussed below in terms of the cognitive processes they targeted. The results for the task evaluation questionnaires are again presented separately for the two groups of test takers.

4.2.3.1 Cognitive requirements of B1 level task according to theoretical frameworks

B1 level task in the first piloting was designed to assess test takers' listening ability "to understand specific information" as in the previous two tasks (See Appendix B for the first B1 level task). For this task, the test takers listened to a conversation between three classmates about the first lesson of a course and its requirements. They were asked to complete the notes about the course based on the listening text. In the CEFR scales, learners at B1 level are considered to be "independent users" of the language as opposed to "basic users" at A1 and A2 levels. Listeners are presumed to

“understand straightforward factual information about common every day or job related topics, identifying both general messages and specific details, provided speech is clearly articulated in a generally familiar accent” and “understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure etc., including short narratives”. We can conclude from these statements in the CEFR specifications that expectations from listeners increase as the level of proficiency increases, which means that questions should go beyond assessing simple, clear and factual information, and extended discussions should become a part of the assessment.

However, the questions in this task all require comprehension of specific information and the only target listening sub-skill is ‘listening for specific information’. The answers required were mostly dates, numbers, a day and a book name. This kind of factual information does not trigger higher-level thinking processes and test takers only focus on specific words to respond. Although there were distractors in the text, it did not encourage test takers to understand the whole text or construct meaning beyond sentence level. However, according to Field’s (2013) listening framework, listeners also need to employ meaning and discourse representation at higher levels. Therefore, a task at B1 level should include items that trigger such cognitive processes. Moreover, when Weir’s listening taxonomy is also taken into account, it can be found that inferred meaning comprehension is the following step in listening comprehension after direct meaning comprehension. Based on these considerations, we can argue that this task failed to fulfill its objectives. When the statistical analyses of this task were carried out, it was observed that this task was easier than the task at A2 level (See section 4.4.1 for a discussion on the scoring validity of the B1 level task in the first administration). As a

consequence, this task underwent changes in terms of question format and language, and later given to test takers as the new A2 level task in the second administration as mentioned in the previous section. Although this B1 level task was expected to pose more difficulty to students in terms of linguistic properties, the topic and the straightforwardness of factual information throughout the text may have lowered this estimated difficulty. It was concluded that the difficulty of the text does not always determine the difficulty of the task, and thus the cognitive skills required to complete the task. Difficulty is actually an inter-play between text, task and expected response characteristics.

As it was stated above, A2 level task in the first piloting became B1 level task in the second piloting with some alterations (See Appendix B for the A2 level task in the first administration and Appendix C for the B1 level task in the second administration). This task was composed of a dialogue between two classmates and some multiple-choice questions. Contrary to what the previous B1 level task measured, this test aimed to assess a wider range of listening sub-skills to suit the target proficiency level more properly. The target sub-skills of this task are listed as below in the test specifications (See Appendix D for the test specifications):

- Listening for specifics, including recall of important details (Weir, 1993)
- Listening for main idea(s) or important information: and distinguishing that from supporting detail, or examples (Weir, 1993)
- Understanding discourse markers (Weir, 1993)
- Identifying and reconstructing topics and coherent structure from ongoing discourse involving two or more speakers (Richards, 1983)
- Determining a speaker's attitude or intention towards a listener or a topic (Weir, 1993)

- Making inferences and deductions at local levels (Weir, 1993)

These target sub-skills cover both lower-level and higher-level processes from Field's (2013) and Weir's (1993) framework. One sub-skill was worded according to Richard's academic listening sub-skills taxonomy; however, it is also in line with Field's framework. Thus, based on these frameworks, it can be stated that the task measures higher-level listening processes. The meaning and discourse construction processes are not stated in the CEFR specifications at B1 level and can be considered as high-level processes; however, at this level listeners are named as "independent users" of the language, and we need to make a distinction between the target sub-skills at A2 and B1 level. Therefore, it is considered appropriate to include some items that require higher-level processes in this task.

4.2.3.2 Cognitive requirements of B1 level task according to task evaluation questionnaires

This task was also taken by two different groups of test takers similar to A2 level task. Thus, the results of the questionnaire are discussed separately for the two groups. In this task, the participants chose from 12 sub-skills in the task evaluation questionnaires for six items to state which sub-skills they used to answer each item in the task (See Appendix E for the task evaluation questionnaires). The results of the questionnaires are demonstrated in Table 4 for the lower-level test takers and Table 5 for the higher-level test takers. The popular choices of the test takers are shown with an asterisk (*) in the tables.

Table 4 shows that the lower-level test takers employed a wider range of sub-skills to respond to the items in B1 level task. The first, third, fourth and sixth sub-skills were utilized the most by the lower-level test takers. This indicates that the test

takers employed higher-level listening processes such as meaning and discourse construction (the fourth and sixth sub-skills). Therefore, it can be argued that these results reflect the theoretical findings mentioned above since the items in this task achieve to measure higher-level listening processes. Moreover, the test takers needed to use the eighth sub-skill to answer the first and fifth items, which required the test takers to use their available information to infer the necessary meaning. These findings show that the items could successfully elicit the target sub-skills.

Table 4. Lower-level Test Takers' Perceptions of Cognitive Processes in B1 Level Task (n=16)

In order to answer this question correctly I had to...	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1. understand specific bits of information in the dialogue	14*	13*	12*	11*	10*	13*
2. understand just the main idea(s)	4	3	0	1	4	4
3. understand the details used to explain the main idea(s)	6*	8*	7*	8*	7*	6*
4. differentiate between important and less important information	10*	8*	8*	8*	8*	6*
5. understand what the dialogue is about briefly	5	3	2	4	3	1
6. understand how information in the whole dialogue fits together	5*	5*	6*	6*	6*	6*
7. pay attention to the speakers' attitude and tone.	5	0	1	2	3	4
8. make an inference based on the information in the text	6*	2	3	3	5*	3
9. understand relations between the speakers and the situation they are in	1	2	3	2	3	2
10. understand what the speaker's intention is when using a certain sentence	2	0	1	2	1	4
11. understand what an unknown word/phrase means based on the information in the text	0	0	0	1	2	1
12. rely on my general world knowledge.	3	4	3	3	4	5

For the higher-level test takers, the results in Table 5 demonstrate that they also adopted similar sub-skills (the first, third and fourth sub-skills) to the lower-level test takers except for the sixth sub-skill; however, the popularity of the sub-skills for each item is not as strong as for the lower-level group. Therefore, we can state that the most popular sub-skill for the higher-level group is the first one, while

there are a few more sub-skills utilized more widely by the lower-level group. This might result from the possibility that the lower-level group lacked the necessary linguistic knowledge to answer the questions and therefore, attempted to compensate for this lack by employing various listening sub-skills. We might also assume that the task could be completed through the use of mostly lower-level sub-skills by the higher-level test takers. However, this should be seen only as a prediction, as we do not have much conclusive data to support this claim.

Table 5. Higher-level Test Takers' Perceptions of Cognitive Processes in B1 Level Task (n=14)

In order to answer this question correctly I had to...	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1. understand specific bits of information in the dialogue	7*	11*	8*	9*	6*	6*
2. understand just the main idea(s)	3	1	3	1	2	3
3. understand the details used to explain the main idea(s)	5*	4*	4*	3	4*	2
4. differentiate between important and less important information	3	7*	4*	4*	4*	3
5. understand what the dialogue is about briefly	2	3	1	3	2	3
6. understand how information in the whole dialogue fits together	3	4	3	3	3	2
7. pay attention to the speakers' attitude and tone.	2	2	2	1	2	3
8. make an inference based on the information in the text	2	1	1	2	4*	3
9. understand relations between the speakers and the situation they are in	1	1	1	2	2	1
10. understand what the speaker's intention is when using a certain sentence	2	0	0	1	1	1
11. understand what an unknown word/phrase means based on the information in the text	1	1	1	2	1	2
12. rely on my general world knowledge.	1	1	1	0	0	1

By looking at the lower-level test takers' results, we can argue for B1 level task that it was able to trigger the necessary listening processes although the results for the higher-level test takers do not imply such a conclusion as strongly as for the lower-level group. Once again, it was noteworthy that different proficiency level test takers reported somewhat different sub-skill use.

4.2.4 B2 level task

B2 level tasks administered in both the first and second pilot administrations are discussed below in terms of the cognitive processes the items in them target.

4.2.4.1 Cognitive requirements of B2 level task according to theoretical frameworks

Some significant changes were also made for B2 level task after the first administration of the test. Due to cognitive concerns, the B2 level task in the first administration was completely omitted from the test and a new test task that would suit the expected cognitive requirements of this proficiency level more was developed by the researcher and the testing expert. Therefore, firstly, the CEFR specifications for this level is mentioned, the problematic areas of the first B2 level task are discussed and the development of the new B2 level task is explained below.

B2 level task in the first administration was designed to measure the test takers' ability 'to listen for specific information' and 'make inferences based on the information in the text' (See Appendix B for the B2 level task in the first pilot administration). Listening ability at B2 level is described as in the following in the CEFR scales: "Can understand the main ideas of propositionally and linguistically complex speech on both concrete and abstract topics delivered in a standard dialect, including technical discussions in his/her field of specialisation" and "Can follow extended speech and complex lines of argument provided the topic is reasonably familiar, and the direction of the talk is sign-posted by explicit markers". According the CEFR, learners at B2 level are considered to be "independent users" of the language. Therefore, the sub-skills targeted should match the proficiency level and therefore demand higher cognitive skills. In Field's (2013) listening framework and Weir's (1993) listening taxonomy, it can be observed that higher-level processes

such as inferences, meaning construction and discourse representation come at later stages of listening comprehension and language proficiency. As a result, this task aimed to go beyond the previous tasks in terms of cognitive requirements as mentioned above.

However, one observation regarding the task type of the first B2 level task was essential and caused a major alteration in the test. For this task, the test takers were supposed to listen to a course announcement and decide if the statements given in the task were 'True', 'False' or 'Not Given' (T/F/NG). This task type, marking sentences T/F/NG did not prove to be a very appropriate method to measure listening ability. Listening is an online activity and listeners need to rely on their working memories briefly to remember the previously mentioned points in the listening. However, the nature of this T/F/NG task type requires the use of memory heavily because of the "Not Given" choice. The problem with this choice is that listeners look at the statements and listen at the same time; however, when one of the statements is not mentioned in the text and the recording continues, listeners may still wait to hear the "Not Given" statement, thus missing the information about the following statements. Since there is no going back in listening unless the recording is played twice, listeners may not listen to the information they have missed again, and may fail to answer some questions because of this task type. Therefore, it was realized after the first administration that "identifying 'Not Given' information" was not a testable listening sub-skill. Deleting the "Not Given" choice from the task could have been one solution to this problem; however, giving test takers 50 per cent chance of guessing the correct answer was not believed to enhance the validity and reliability claims of the test. Especially at a higher-level task such as this one, comprehension of test takers should be tested more carefully and in a more detailed

way. As a consequence, this task was discarded from the test and replaced by an entirely new one.

Another problem with this task was the content of the text. The issues related to content are discussed in the investigation of the second research question; however, as previously stated in Chapter 2, Weir argues that theory-based, context and scoring validity types are closely related to one another and they have implications regarding the other validity types. In this task, the content of the listening text also affects the cognitive requirements of the task as well; therefore, it needs to be discussed in this section, too. Since the listening script was based on a course announcement, the text was generally composed of factual information, except for some parts where the speaker talked about her opinions regarding the course. Since it mostly contained factual information such as the purpose of the course, time, price, participants, etc., it did not assess understanding of more complex meaning relations such as discussions and agreement/disagreement in abstract topics. This resulted in asking questions about specific details, and explicit and factual information, which did not satisfy the cognitive requirements of this proficiency level. It was concluded that a course announcement does not provide much opportunity for this kind of contexts and may not be a very appropriate text type to measure B2 level listening proficiency.

Moreover, although the linguistic difficulty of the text was considered to be at B2 level because of its wide variety of complex structures and a wide range of both frequent and less frequent vocabulary items, the text difficulty in terms of meaning sophistication was not at the expected level because of the topics included in the listening script. Thus, it was concluded that syntactic and lexical difficulty does not always match and guarantee the cognitive difficulty of the text itself. Due to all these

reasons, it was decided after the first piloting that the task had to be changed altogether in order to develop another one that is more appropriate for the cognitive expectations at this level.

After the omission of the B2 level task in the first administration, a new one that assesses the expected cognitive processes more appropriately was developed by the test writers. The new task was composed of eight multiple-choice questions and the test takers were supposed to listen to a radio program about a festival (See Appendix C for the B2 level task in the second administration). The target sub-skills were also chosen from a variety of lower-level and higher-level sub-skills according to Field (2013), Weir (1993) and Richards (1983). The sub-skills aimed in this task are as follows (See Appendix D for the test specifications):

- Listening for specifics, including recall of important details (Weir, 1993)
- Listening for main idea(s) or important information: and distinguishing that from supporting detail, or examples (Weir, 1993)
- Identifying role of discourse markers in signaling structure of a text (conjunctions, adverbs, etc.) (Richards, 1983)
- Identifying and reconstructing topics and coherent structure from ongoing discourse involving two or more speakers (Richards, 1983)
- Making inferences and deductions at both local and global levels (Weir, 1993)
- Determining a speaker's attitude or intention towards a listener or a topic (Weir, 1993)

The items in the new B2 level task were designed to assess the sub-skills mentioned above. There are items that require sentence-level factual information, meaning representation and discourse representation. Phonological and lexical information was targeted indirectly in the task. This task aims to cover all the aspects

of Field's (2013) framework since this task has the highest level of proficiency in the whole test and B2 level should cover high-level comprehension skills provided that contextual characteristics are suitable to the level. In Weir's (1993) terminology, both direct and inferred meaning comprehension skills are measured by the items in the task. Thus, the items and the cognitive requirements of the task comply with the expected listening sub-skills in the CEFR scales given above. When these are taken into account, this task is deemed to be a significantly improved measure of listening skill at B2 level compared to the previous version.

4.2.4.2 Cognitive requirements of B2 level task according to task evaluation questionnaires

The task evaluation questionnaire at B2 level was given for the new B2 level task in the second administration. In this task, the test takers were asked to answer eight multiple-choice questions and then select from 12 sub-skills in the questionnaires to indicate the sub-skills they used while answering the specific items. This task was delivered to only the higher-level test takers; therefore, only the results for this group of participants are given below. Table 6 outlines the results for B2 level task. The important findings are demonstrated with an asterisk (*) in the table.

It can be seen in Table 6 that the most chosen skills vary in this task, as the target skills also differ. Most of the higher-level test takers heavily employed the first, third, fourth and eighth sub-skills for almost all of the items and adopted the fifth and ninth sub-skills for some of the items in the task. This provides evidence for the level of variation across tasks in terms of the cognitive processes required. As the most complex task, B2 level task seems to trigger a combination of lower and higher-level listening comprehension skills. The test takers needed to understand specific

information, distinguish between different types of information, make inferences, and form meaning and discourse representation in their minds to be able to answer the questions. Most of the questions aim to assess comprehension of the implied meaning of the utterances rather than their literal meanings. This is especially assessed by paraphrasing the items and the options especially at points where answers could be found. Consequently, the test takers had to construct meaning and discourse in their minds by paying attention to the details and the importance of the incoming information. This is why the popular sub-skills are the first and fourth ones for the first and sixth items.

Table 6. Higher-level Test Takers' Perceptions of Cognitive Processes in B2 Level Task (n=14)

In order to answer this question correctly I had to...	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
1. understand specific bits of information in the dialogue	8*	7*	8*	7*	5*	5*	5*	3
2. understand just the main idea(s)	3	3	2	2	3	2	2	2
3. understand the details used to explain the main idea(s)	5*	5*	5*	5*	4	4	6*	7*
4. differentiate between important and less important information	6*	5*	8*	7*	7*	6*	6*	5*
5. understand what the dialogue is about briefly	5*	3	4	4	2	2	8*	5*
6. pay attention to the speakers' attitude and tone	1	1	1	1	1	1	3	1
7. understand how information in the whole dialogue fits together	1	1	1	2	2	3	2	7
8. understand how certain parts are linked to others in the dialogue	5*	4	5*	5*	6*	6*	6*	8*
9. make an inference based on the information in the text	2	2	3	3	5*	6*	4	7*
10. understand what the speaker's intention is when using a certain sentence	0	0	0	0	1	1	1	0
11. understand what an unknown word/phrase means based on the information in the text	2	1	2	1	1	2	2	1
12. rely on my general world knowledge.	3	3	2	2	2	2	2	2

However, for especially the seventh and eighth items, which differentiate the task from the others in terms of the target processes, the most widely chosen sub-

skills (the fifth, eighth and ninth sub-skills) differ compared to items one to six. The seventh and eighth items tap into more global levels of processing and therefore, the test takers had to combine the main ideas and certain parts in the text to be able to answer the questions. For the seventh item, the test takers were requested to determine the social purpose of the festival mentioned in the listening. Although the answer to this item is stated in the last part of the dialogue, it is also essential to combine it with the information throughout the text. According to the results of the task questionnaire, the item seems to have achieved its purpose and elicited the necessary cognitive processes specified in the third, fourth, fifth and eighth sub-skills. Similarly, the eighth item requires parallel processes to the seventh item; but it refers to a more global inference regarding the whole listening text. Therefore, test takers needed to combine all the information they had heard and make an inference based on it. This is indicated in the instructions right before the item. It was stated that the question would be answered after the listening finished. The results of the task evaluation questionnaire indicate that this item also succeeds in stimulating the target cognitive processes since the third, seventh, eighth and ninth sub-skills were chosen by a majority of the test takers. Overall, for B2 level task, we can conclude that it succeeds in eliciting a wide variety of sub-skills, which were stated in the test specifications.

4.2.5 C1 level task

C1 level task was only administered in the first piloting and it was not included in the second administration of the test. The reasons for excluding this task from the test are mentioned below, but no modifications are explained, as it was not modified but directly deleted from the test.

In this task, the test takers were expected to listen to an extract from an authentic lecture and fill in the gaps in the notes about the lecture (See Appendix B for C1 level task). Since this test was mainly developed as an academic listening test, it was considered appropriate to include a task which required listening to a lecture in the test because lectures are a major component of academic life and listening to a lecture is an essential academic skill. Richards (1983) provides a comprehensive list of academic listening sub-skills which should be integrated into a task that measures academic listening skills and we tried to operationalize these sub-skills in C1 level task. However, although this task was designed to assess academic listening sub-skills, some problems regarding cognitive validity were raised by the TFL instructors and the testing expert after the first piloting. The subject matter in the listening text was an introduction to computers and the ways computers are connected to each other and data are transferred. The lecture was an authentic material from distance education materials. The topic was intentionally chosen since the test writers did not want test takers to rely on their background knowledge much in order to answer the questions. Therefore, a lecture with some new and different terminology was chosen. Nevertheless, the problem with the lecture chosen for this task was not the predictability of the answers, but the amount of the factual information it contained. Since the lecture was mostly on terminology and explanations, it did not include much discussion in adequate depth to be appropriate for higher-level comprehension questions. At this proficiency level where learners can understand extended speech and discussions without difficulty, asking questions about specific details did not satisfactorily meet the cognitive requirements of the expected level. TFL lecturers also did not find the task relevant to the level of the students. Therefore, this task was considered to fail to measure the target sub-skills at C1 level.

This finding again demonstrates that the genre of the text does not always guarantee a certain level of cognitive demands. Including a lecture in a test may not always provide us with satisfying results. Not only the topic or the general genre of a text but also rhetorical purpose and organization of information in it are also important characteristics that should be taken into consideration as designated in Weir's (2005) framework. The lecture chosen should stimulate higher-level listening processes and contain intricate lines of argument to prove to be an appropriate way of assessing the target proficiency level. As a result of all these observations and findings, the task was discarded from the test. In the second piloting, no new tasks including a lecture were added. Since B2 level is a level where students can function well at universities, no C1 level task was developed in the second piloting and the new version of the test was composed of four tasks ranging from A1 to B2 levels of proficiency. Since a task at C1 level was not delivered in the second administration of the test, no task evaluation forms were given to the test takers; therefore, there is no discussion on the results of a questionnaire for C1 level task.

4.2.6 Conclusion for the investigation of theory-based validity

In conclusion, we can argue that the cognitive requirements of different proficiency levels are graded and varied across tasks based on the argumentation provided for this research question. As stated above, some modifications needs to be done regarding A2 level task to distinguish it from A1 level task and items that target sentence-level factual information can be integrated into A2 level task. The results of the task evaluation questionnaires also indicate that the targeted cognitive skills are generally elicited by the items in the tasks. Moreover, it can be observed in the results of the task evaluation questionnaires that the test takers tended to choose from

the sub-skills from the lower parts of the questionnaires gradually with the increasing proficiency levels. The sub-skills at the top of the list in the questionnaires are relevant to more local sub-skills, whereas those at the end of the list refer to more global ones. Thus, based on the findings, it can be stated that the variation across the tasks in terms of theoretical requirements can also be seen in the task questionnaires. Another important finding is that the lower-level test takers tended to choose more diverse sub-skills in the task evaluation questionnaires compared to the higher-level test takers. They might have had to employ various listening sub-skills in order to answer the items and compensate for their lower proficiency level. As discussed before, this indicates that test takers at different levels of proficiency can employ different sub-skills to respond to the same items.

4.3 Investigation of context validity

In the previous part, the cognitive processes involved in the listening process and the way the tasks in the test can tap into these processes were taken into consideration. In this part of the study, the focus will shift from the cognitive processes to the contextual issues about the test such as task setting, administration setting and linguistic demands of tasks and speakers. The second research question, which has two sub-questions, was created to scrutinize these issues related to context validity in the current test. The research questions can be seen below:

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

- a. What are the demands imposed upon the test takers by task setting, administration setting, linguistic features of the listening test tasks and the speakers?

- b. What are the participants' perceptions of the tasks in terms of the suitability of their contextual features for the different proficiency levels?

In the following pages, the tasks in the present study are scrutinized by taking into account the components of context validity as outlined by Weir (2005) in Figure 2 in section 2.3 and the theoretical discussions provided by Weir (2005) and Elliott and Wilson (2013) on these components.

4.3.1 Task setting

4.3.1.1 Purpose

The purpose of the tasks is closely related with authenticity as discussed in section 2.3.3.1. In the study under investigation, most of the listening texts used are not authentic but scripted. The listening text for C1 level task in the first piloting was a fully authentic recording of a lecture. In addition, the B2 level listening text in the second administration was a semi-scripted text. For this text, an authentic radio program was modified to make it more suitable for a language test. Apart from these two, the texts for the other tasks both in the first and second administration were written by the researcher and the testing expert. One reason for this was the scarcity of authentic listening materials especially for lower-level proficiency levels in Turkish. Moreover, Richards (2007) discusses the unrealistic goals for using authentic materials because of many difficulties such as their linguistic features, difficulties in recording, copyright issues and consent of the participants. In addition, adapting the texts can be even more difficult for some tasks, especially multiple-choice tasks; therefore, the test developers chose to write the scripts themselves.

The purpose of the tasks is also concerned with the type of tasks included in a test and the level of focus that the tasks require from the test takers as stated in section 2.3.3.1. The task types in the current test are short-answer questions for A1 and A2 levels, and multiple-choice questions for B1 and B2 levels (See Appendix C for the tasks in the second pilot administration). According to Figure 3 given in section 2.3.3.1, it can be stated that the tasks in the present test mostly taps into medium, deep and very deep attentional foci with both global and local aspects. Both short-answer and multiple-choice questions are eligible for eliciting these types of focus and a variety of different listening sub-skills. Skimming and unfocused scanning are not assessed in this test, as the number of tasks is limited.

The rubric, or the instructions, given to test takers in an examination is another consideration in terms of task purpose. In the current test, the instructions are given in both written and spoken forms to the test takers before each recording. It is composed of the instructions regarding the setting, the topic, the speakers, the information about how to complete the tasks and the allocated time to read the items. When there are different test formats in one task, they are also specifically stated. One example comes from A2 level task where there are multiple-choice, gap-filling and short-answer questions in the same task. Since test takers are supposed to write the answers in a blank for both gap-filling and short-answer questions, the same instruction is given for them. However, for the MCQ format, students are also told to choose the correct answer in the instructions. The instructions for this task are demonstrated below:

“Şimdi iki arkadaş arasında yurttan geçen bir konuşmayı dinleyeceksiniz. Konuşmayı bir kere dinleyeceksiniz. Önce soruları 1 dakika içinde okuyunuz. Doğru cevapları işaretleyiniz ya da boşluklara yazınız.” (Now you are going to listen to a dialogue between two friends at a dormitory. You are going to listen to it once. Firstly read the questions in 1 minute. Choose the correct answer or write the correct answers in the blanks.)

Another case where a different instruction is given is for the 8th item in the B2 level task in the second administration. This question should be answered after the listening has finished, since it requires the test takers to synthesize all the information in the listening text; therefore, right before item 8, a separate instruction was given as in the following:

“8. soruyu dinleme bittikten sonra cevaplayınız.” (Please answer the 8th question after the listening has finished.)

Another issue about giving instructions for different task types is the multiple-choice questions whose stems are not similar to one another (Elliott & Wilson, 2013). In B1 level task in the second piloting, although all of the questions are in multiple-choice format, the item stems differ from each other, which makes testing focus different. For the first four questions, test takers need to choose the best sentences that answer the questions, while they need to choose the best sentences that complete the item stems in the fifth and sixth questions. For these kinds of differences, no extra instruction is given. In this case, test takers are supposed to read the items to be able to comprehend the purpose of the task and the type of information it requires.

4.3.1.2 Response format

The response methods in the current test are analyzed according to the considerations outlined by Elliott and Wilson (2013) in section 2.3.3.1. They maintain that the rubrics, item stems, note-taking procedures during the exam and the memory load on the test takers need to be considered carefully. The rubric in this study has already been discussed above and the kind of information it contained has been stated. As mentioned in section 3.5.2, in the third section of the task evaluation questionnaires,

the test takers were asked to evaluate the contextual characteristics of the tasks and one of these was related to how clear the instructions were (See Appendix E for the task evaluation questionnaires). The test takers were asked to agree or disagree with the statement “The instructions were clear” on a five-point scale from 1 (*definitely agree*) to 5 (*definitely disagree*). Their evaluation of the clarity of the instructions is given below in Table 7 separately for lower-level and higher-level test takers, as tasks A1 and B2 were not delivered to both groups. The table shows that the test takers agreed that the instructions were clear, as all of the mean scores are between “1” and “2”. In brief, the information in the rubric was considered to be sufficient and clear by the test takers.

Table 7. Test Takers’ Perceptions of the Clarity of the Instructions

Task	Mean scores for lower-level test takers	Mean scores for higher-level test takers
A1 Level Task	1.6	-
A2 Level Task	1.3	1.2
B1 Level Task	1.2	1.3
B2 Level Task	-	1.07

Elliott and Wilson (2013) discuss that preparing the answer key also needs to be taken into account while choosing the response format. The scoring for the current test is considered to be objective, especially for the multiple-choice tasks. In the first two tasks, the test takers are supposed to write very short answers, one-word answers almost all the time. As a consequence, marking them was quite objective, as the items did not yield any complicated answers. While marking the tests, when logical answers appeared for some questions, they were evaluated and included in the answer key after the piloting sessions. The multiple-choice questions had only one standard answer and therefore, they were quite easy and objective to mark.

The linguistic difficulty of the item stems is another important consideration in selecting response format (Elliott & Wilson, 2013). If test takers cannot answer a

question because of its linguistic difficulty, this may lead to construct-irrelevant variance. Therefore, in this test, the items were written with easier vocabulary and grammatical structures compared to the listening scripts. However, after the second administration, one suggestion was received from the higher-level test takers about B2 level task. They reported that they did not understand the words “öğē” and “geçit” in the items. Although “geçit” is a word mentioned in the text a few times, it caused difficulty to the test takers. The other word, “öğē”, was in the correct option, and the test takers found this word challenging, too; as a result, it may have caused high item difficulty. In the Turkish National Corpus (Aksan et al., 2012), these words are shown to be low frequency words in the word frequency lists; “geçit” is the 9056th word in the list with an observed frequency of 588, and “öğē(leri)” is the 8968th word with an observed frequency of 594. Low frequency words such as these should be investigated properly before they are used in a test. In this respect, using corpus for determining the linguistic difficulty of listening texts can be considered helpful.

In terms of note-taking, no specific instructions were given and no restrictions were applied. The test takers were not specifically told to take notes during listening, but they were not forbidden, either. One reason for this is that in the present test, the questions are previewed before the listening; therefore, their note-taking skills are not measured in any of the tasks. As a result, it was not deemed necessary to adopt a policy towards note-taking. If a new task requiring note-taking is included in the test in the future, instructions about note-taking will then be integrated into the test specifications and the rubric.

Memory load also needs to be controlled in a listening test. The memory load imposed upon the test takers by the test tasks in the current study can be considered

to be low at A1 and A2 levels, as they only wrote one or two words, or numbers as answers. However, since B1 and B2 level tasks are in multiple-choice format, the memory load is relatively higher. In such a response format, test takers need to bear in mind the options, the incoming information, and the previously stated information to build meaning representation and make a decision about the options. Because of this response format, test takers need to rely on their memory relatively more than in other response formats. After the second piloting, one suggestion from the experts regarding multiple-choice format was to order the multiple-choice options according to the place the relevant information appears in the listening text to make it easier for test takers to follow the options and to decrease the memory load. This suggestion was valuable because a complex task such as listening multiple-choice, where test takers need to carry out many tasks at the same time, should be made more dependent on listening text comprehension and less dependent on memory, reading and other construct-irrelevant variances. This revision will be made in the future versions of the test.

In addition to the important parameters about response formats, the specific the task types in a listening test should be investigated carefully as well. Based on Khalifa and Weir (2009) as discussed in section 2.3.3.1, the test in the present study includes both constructed and selected response formats; short-answer questions as constructed response format and multiple-choice questions as selected response format. The pros and cons of using these response formats in a listening test are discussed elaborately in section 2.3.3.1 and will not be mentioned here again to avoid repetition. It should be stated that the positive and negative features of these task types were taken into consideration during the development of the current test.

In the MCQs, in order to prevent test-wiseness strategies, the distractors in the items are all included in the recorded text so that test takers need to listen for all the related information rather than simply matching the words in the item and the text. They will not only hear and identify the correct option, but will also disconfirm the others. As stated before, this situation may cause concerns about cognitive validity since this is not how listening occurs in real life. However, despite this disadvantage, reliability, ease of marking and flexibility in terms of tapping into various levels of processing make MCQs desirable in listening tests (Elliott & Wilson, 2013) and this response format will be kept in the current test. As to the number of options in MCQs, both sets of MCQs in this test have four options. Considering the findings put forward by Rodriguez (2005, in Elliott & Wilson, 2013), Moreno, Martinez and Muñiz (2006, in Elliott & Wilson, 2013) and Boroughs (2003, in Elliott & Wilson, 2013) as discussed in section 2.3.3.1, we can suggest that MCQs at B1 level could have three options, while those at B2 level could have four options to make them more suitable for the target levels.

With respect to the constructed response format (short-answer questions), some issues regarding spelling, the answer key and the length of the answers are important parameters that need to be taken into account (Elliott & Wilson, 2013). In the current test, the spelling policy is indicated in the test specifications. As discussed in section 2.3.3.1, a limited range of spelling mistakes are accepted in this study. Some accepted and unaccepted answers from the test with regard to spelling are shown in Table 8. As it can be seen from Table 8, in constructed responses only one change in sound/letter is accepted as long as the word is not totally changed into another one. In this test, learners are supposed to write very short words, which are generally high-frequency words that are supposed to be within their capacity and

numbers. Thus, allowing only minor mistakes was feasible. Furthermore, Turkish is a transparent language, which has sound-symbol correspondence. This is supposed to help test takers and they are expected to hear the words and write them as heard. This policy is adopted considering the CEFR specifications: At A1 level, learners “can write simple isolated phrases or sentences.” At A2 level, they “can write a series of simple phrases and sentences linked with simple connectors like ‘and’, ‘but’ and ‘because’”. Since the tasks are in line with these sub-skills, test takers should be able to carry out the tasks.

Table 8. Sample Responses and Results for Spelling Mistakes

Task Level	Test taker's responses	Correct answers	Results
A1	ilaj	ilaç	Accepted
A1	sonum	sunum	Accepted
A2	bala	bahar	Not accepted
A2	buhur	bahar	Not accepted
A2	çarşamaba	çarşamba	Accepted
A2	makare	makale	Accepted
A2	sorum	sunum	Not accepted

In terms of the clarity of the key, in the test under study, the key is mostly strict, since the answers do not yield any other possible responses. In multiple-word answers such as “tükenmez kalem” in the first item of A1 level task, the most essential information “kalem” is also regarded as a correct answer. Similarly in the same task, the answer “ağrı kesici ilaç” in the sixth item can be also accepted as “ağrı kesici” or “ilaç” since they also refer to the same thing and there are no other “ilaç” types mentioned in the text. There are also some numerical answers in the text and these are, of course, accepted in both numerical and lexical forms, i.e. “2” or “iki”.

Elliott and Wilson (2013) also mention that the length of the short answers in constructed response formats should generally be specified. In this test, the number

of words for the answers is not specifically stated. Therefore, in the following versions of the test some modifications can be done regarding this issue.

In conclusion, in the present test, most crucial aspects of response format have been taken care of. The minor suggestions made above can be helpful in increasing further the theory-based and context validity of the test formats.

4.3.1.3 Known criteria

As mentioned in section 2.3.3.1, test takers need to know the criteria that they are being assessed against (Elliott & Wilson, 2013). In this test, the test takers were not provided with any information about the marking criteria. The reasons for this are that the administrations were only pilot studies and the test takers knew the test scores were not a part of their assessment. Therefore, they did not need to worry about the test results. In a real test administration, marking criteria needs to be included in the instructions. Another aspect that needs to be modified is the criteria for spelling. The policy for spelling mistakes should be indicated in the rubric, too.

4.3.1.4 Weighting

In this test, there are no weighted items and all are assigned the same mark. In the test specifications, it is stated that all the items are scored dichotomously, either “1” or “0”. Therefore, the test takers were not given any information about weighting.

4.3.1.5 Order of items

As discussed in section 2.3.3.1, listening is an event that happens online and listeners cannot listen to the same text again; therefore, the order of the items in a listening test is crucial. In the listening test in this study, the items are in a linear order; that is

to say, they appear in the order that the information appears in the text. As stated before, one suggestion regarding B2 level task after the second piloting was to order the options in the MCQs as well according to the information in the text as well in order to decrease the reading load on test takers. When items are in an order, test takers will be able to follow the ideas more easily and will not have to go back and forth between options at the same time. This will help reduce the cognitive and memory load on test takers.

As pointed out by Elliott and Wilson (2013), time spaces between items should also be carefully considered in a listening test. In Table 9, you can see the ratio of timings to items across different tasks in the second administration of the study under investigation. The results in Table 9 are calculated by dividing the total recording lengths in seconds by the number of items in each task.

Table 9. Ratio of Timings to Items across Tasks

Task	Ratio of timing
A1 Level Task	25 seconds
A2 Level Task	21 seconds
B1 Level Task	29 seconds
B2 Level Task	55 seconds

As it can be seen from Table 9, there is a gradual increase among the tasks except for A2 level task. This may result from the nature of the task. It is a gap-filling task with a factual listening text, which requires one-word answers. Therefore, the gaps between the answers do not need to be very long. In addition, there are more items in A2 level task than in A1 level task. However, the amount of time for each item can still be regarded as enough depending on the observations from the pilot administrations. The fact that the spaces between items in this test increase with the proficiency levels can be explained by the amount of information that test takers need to process from the text and the items. In relation to this, a dramatic increase

from task B1 to task B2 can be observed. This can be explained by the cognitive demands of the items. B1 and B2 level tasks are multiple-choice tasks with four options and all the answers and distractors are integrated into the text. However, due to the linguistic difficulty of the B2 level listening script and the more global cognitive processes targeted, B2 level task required more time for the items. The test takers also needed to read the relatively longer options while taking the task; therefore, leaving longer spaces between items seemed necessary. One piece of feedback was received from the Turkish instructors before the second piloting with respect to the ratio of timings to items. It was reported that some parts of the B2 level listening text was too long and they did not contain any answers to the items. It was mentioned that listening to long passages in the text by searching for an answer but not finding it might be misleading to the test takers. Having longer recordings and few items in tests is also mentioned by Elliott and & Wilson (2013) as a problem. As a result of this feedback, some revisions were made in the script and some redundant parts were omitted or transformed into distractors.

In conclusion, order of the items, time intervals between items and item density were taken into account in the development of the present study, and operationalized in both administrations of the test. These criteria should be given careful consideration and testing the tasks with a few people, e.g. some language learners, native speakers or experts, can help notice problems related to them before the real administration of the test as was the case in the present study.

4.3.1.6 Time constraints

How much time will be allocated to test takers between two listening tasks and between two items in a task is another consideration about task setting (Elliott &

Wilson, 2013). In the current test, different recordings were prepared for each task. Each recording consisted of three components: the instructions in spoken form, the allocated time for the items to be read recorded in silence, and the recorded text. Test administrators only needed to press the play button for each task. From A1 to B1 levels, after the candidates listened to the instructions, they were given one minute to read the items, and the options if there is any. After one minute, the recording started and they answered the questions. However, for the B2 level task, they were given three minutes to go over the questions after hearing the instructions. The reason for this is the reading load imposed by the task format. They needed to read eight questions, each of which had four options in mostly sentence length. Therefore, it is considered appropriate to allow a longer period of time for the test takers to read the items in more challenging tasks. After the listening, the test takers were given two minutes to complete answering the questions in each task. As mentioned previously, since this is a validation study, the test takers were also asked to complete a questionnaire regarding the cognitive and contextual requirements of the tasks after each task finished. They were given nearly five minutes to complete each questionnaire and then the new recording and task started.

The recorded texts are played only once in the current test based on the premise that in a real-world context listeners cannot listen to the same speech more than once. Instead of playing the texts for the second time, parameters such as the clarity of speech, redundancy of information, time spaces between items and so on, were controlled to make the speech processable at the designated levels. Therefore, in the present test, recordings are heard only once due to concerns regarding cognitive validity (in terms of the nature of spoken interaction).

The test takers also stated their opinions regarding the times the listening is heard in the task evaluation questionnaires (See Appendix E for the task questionnaires). In the third section of the questionnaires, the test takers evaluated the following statement: “It was enough to listen to the text once”. They marked a five-point scale ranging from “*definitely agree* (1)” to “*definitely disagree* (5)”. The results of the questionnaire can be seen in Table 10. The mean scores in Table 10 demonstrate that the lower-level test takers were neutral about the A1 level task, had a slightly positive opinion on the A2 level task and had a slightly negative opinion about the B1 level task. The lower-level test takers might have felt the need to listen to the texts more than once; however, the results do not make very strong implications for this. One observation about the results is that the test takers needed to listen to A1 level text more than A2 level text as the higher mean score for A1 level text shows. Therefore, A1 level task needs to be analyzed in terms of textual and item characteristics. Another important finding is that as opposed to the lower-level test takers, for A2 level task, the higher-level test takers were satisfied with the time of playing the recording and for B1 and B2 levels they were slightly on the positive side. The fact that different groups of test takers showed differing views on this statement means that the difficulty levels of the texts were well determined and that the texts can differentiate between higher and lower-level test takers. Considering these findings, single-play policy can be kept in the present test both because of cognitive validity concerns and the evaluations of the test takers.

Table 10. Test Takers’ Perceptions of the Times Recordings are Heard

Task	Mean scores for lower-level test takers	Mean scores for higher-level test takers
A1 Level Task	3.03	-
A2 Level Task	2.5	1.53
B1 Level Task	3.34	2.38
B2 Level Task	-	2.5

4.3.2 Administration setting

Test administration forms a part of Weir's (2005) socio-cognitive framework since it also has an impact test performance. Since the administrations conducted as a part of this study were only pilot studies, the necessary precautions were taken but policies regarding administration setting are not stated in the test specifications. The tests were delivered at Boğaziçi University by the class instructors, who had been informed about the testing procedure beforehand. If this test becomes an institutionalized test, then procedures for test administration will need to be set with the other parties involved in delivering the test.

4.3.3 Task demands (Linguistic)

Next, the linguistic demands imposed by both the input and the output of the tasks are investigated with relation to the test under investigation.

4.3.3.1 Discourse mode

Discourse mode of the texts and text purpose influence the linguistic demands imposed upon the test takers. Based on Figure 4 in section 2.3.3.3, the test in the present study can be analyzed in terms of text purpose and discourse modes included in the listening texts. The analysis of the texts is shown in Table 11. The table indicates that at the lower proficiency levels, discourse modes are mostly related with the personal surroundings of the people or the topics in their immediate relevance (expressive (of individual)). In addition to expressive (of individual) discourse mode, exploratory discourse mode can be observed as well in A2 and B1 texts that include dialogues with personal opinions, questions and answers, and solutions to problems. In B2 text, the discourse mode is not about personal matters, but on external issues

happening in the outside world. The speakers have a dialogue about a festival and the issues related to it in an interview in B2 text. Therefore, both informative and exploratory discourse modes can be found in this text. These findings indicate that the texts in the test seem to have a variety of text purposes.

Table 11. Text Purpose and Discourse Modes Across Tasks

Tasks	Purpose
A1 level task	Expressive (of individual)
A2 level task	Exploratory / Expressive (of individual)
B1 level task	Exploratory / Expressive (of individual)
B2 level task	Informative / Exploratory

4.3.3.2 Channel of presentation

In the current test, two main channels of presentation are used; i.e. written and spoken, and other alternatives such as pictures, visuals or videos are not utilized. Especially the quality of the audio should be given paramount importance since the main source of input is provided via the audio. In this test, recordings are played on loudspeakers, not on personal headphones. However, the audios are played in a classroom atmosphere where there were limited sources of distractors and noise. In addition, the sound quality of the audios was checked before listening and it was found to be of sufficient quality by the researcher, the testing expert and the TFL instructors.

4.3.3.3 Text length

The level of proficiency is important in determining text length. According to the CEFR descriptors, until B2 level of proficiency, the text length is not described as long, although what “long” means is not specified (Elliott & Wilson, 2013). At B2 level, listeners are described as being able follow extended pieces of speech. This

shows that before B2 level the texts should not be too long and there should be a difference between the B2 level texts and the others.

The text lengths and the total number of words in texts in the present test are given in Table 12. When the data are analyzed, it can be seen that there are variations across levels. The total number of words in A1 level task is more than that of A2 task; however, it is not considered as a problem since there are three short dialogues in A1 level task and only one longer dialogue in A2 task. Therefore, some phatic words and phrases for thanking, greeting and taking leave are repeated in three different contexts in A1 text. Because of these, it is normal that words in A1 level task are more than those in A2 level task. On the other hand, in terms of text length and delivery speed, we can see that A1 and A2 texts are not differentiated enough. This can be taken care of either by slowing down the delivery or shortening the text in A1 level task.

Table 12. Text Lengths across Different Tasks

Tasks	Total length	Total number of words in texts
A1 Level Task	4 minutes in total	356 with instructions
	2 minutes 29 seconds (only the spoken text)	336 without instructions
A2 Level Task	4 minutes 7 seconds in total	337 with instructions
	2 minutes 49 seconds (only the spoken text)	311 without instructions
B1 Level Task	4 minutes 10 seconds in total	418 with instructions
	2 minutes 53 seconds (only the spoken text)	389 without instructions
B2 Level Task	10 minutes 54 seconds in total	859 with instructions
	7 minutes 21 seconds (only the spoken text)	835 without instructions

Another issue that needs to be explained here is the relatively higher number of words in B2 level task. The situation in this task may result from the content and linguistic load of the recorded text. At B2 level, listeners can follow long speeches according to the CEFR descriptors; however, such a dramatic increase may raise some concerns. A sharp increase in text length could cause fatigue and lack of attention and therefore, this problem needs to be mitigated. One revision with respect

to this text could be to shorten it since it is nearly twice longer than the B1 text. However, the text should still be a long one because the level of proficiency and cognitive processes necessary at this level require extended speech. Another solution could be to decrease the number of items in the task. Due to the nature of the task and the number of options, the text is loaded with information. When some of the items are omitted, the related information can also be deleted and this will make the text shorter. These items can be chosen from those that measure word or sentence level factual information. However, in such a case, item quality measures should also be taken into consideration.

4.3.3.4 Nature of information

The information presented in texts can be concrete or abstract as pointed out in section 2.3.3.3. In the CEFR descriptors, at A1, A2 and B1 learners can understand concrete words while they can understand both concrete and abstract words at B2, C1 and C2 levels. The nature of information for each level is indicated in the test specifications of the test (See Appendix D for the test specifications). In the current test, at A1 and A2 levels only concrete words are used. At B1 level mostly concrete words are used, but certain frequently used abstract words are also used, i.e.

“özmek” or “aklından çıkmak”. At B2 level, a mix of concrete and abstract words are included in the text. “gurur”, “emek”, “çıkış noktası”, “etkili”, “büyüleyici”, “katkı”, “heyecanlı”, “masalsı”, “yaratıcı”, “kaynaştırıcı”, and “kutuplaşma” can be shown as some examples of abstract words in the B2 text. In addition to these, there are a number of concrete words which are related to festivals, festival organization, stories, story types, participants, events, and festival venue. Therefore, it can be

argued that the tasks in the current test contain a variety of both concrete and abstract words which are aligned with the target proficiency levels.

4.3.3.5 Content knowledge

In the present test, no background or subject knowledge is necessary to be able to answer the questions, as all answers are embedded in the text. The kind of information presented in the text is also parallel to the kind of knowledge learners have at certain proficiency levels. Table 13 shows the kind of topics that learners are familiar with at certain proficiency levels according to the CEFR descriptors and the topics and the sub-topics included in the current test. Since this test is aimed to measure academic, or in lower levels academically related listening skills, the immediately relevant areas are considered to be school environment and related places. The topics of the texts in this test were chosen from educational domain in the CEFR, which includes a variety of topics that learners are likely to encounter in an academic setting.

Table 13. Topics in the Texts and Their Compatibility with the CEFR Descriptors

Task	Topics / Sub-topics	Corresponding CEFR descriptors
A1	- Shopping stationery goods, their prices, types - Student documents, and necessary processes - Health problems, doctor's examination and recommendations	Areas of immediate need or very familiar topics
A2	- A minor health problem - Introduction to a course and its requirements such as exams, exam dates, assignments, course time and place	Areas of most immediate relevance
B1	- Leisure activities of students, student clubs, family events and relationships - Course requirements such as exams and suggestions about a course-related problem	Familiar matters regularly encountered in work, school, leisure, etc.
B2	- Information about a story festival including its time, place, activities, participants and story - The objectives of the festival and its benefits to the society	Concrete and abstract topics, including technical discussions in his/her field of specialisation

As it can be understood from the information presented in Table 13, no previous knowledge or special subject knowledge is required to be able to respond correctly to the items. On the contrary, the topics chosen resemble those that students in an academic context encounter very often. In addition, no culturally biased information is included in the texts so as not to put anyone at a disadvantage. Although it is indicated in the CEFR descriptors for B2 level that speakers at B2 level can understand topics related with their field of specialization, the B2 text chosen for the present study does not require knowledge of any special field. The topic “festival” was chosen as a semi-scientific dialogue, which can be included in the field of general “social sciences”. Thus, it was attempted to avoid favoring any test takers with any specific background and subject knowledge. As a result, we can say that the topics and the possible existing knowledge of test takers seem to match in the current test.

We also need to support these claims with the test takers’ opinions collected through the task evaluation questionnaires where they evaluated the following sentence about text relevance on a five-point likert-scale ranging from 1 (*definitely agree*) to 5 (*definitely disagree*): “The text was relevant to what I listen to in real life”. The results of the questionnaire are shown in Table 14. The mean scores show that the test takers generally agreed with this statement strongly. Thus, it can be concluded that the test takers were familiar with the topics in the test, which also supports the theoretical discussion given above.

Table 14. Test Takers’ Perceptions of the Recorded Texts in terms of Relevance

Task	Mean scores for lower-level test takers	Mean scores for higher-level test takers
A1 Level Task	1.87	-
A2 Level Task	1.62	2
B1 Level Task	1.75	1.76
B2 Level Task	-	1.64

4.3.3.6 Lexical Resources

In the preparation of the current test and tasks, the lists for vocabulary items in the Reference Level Descriptors as explained in section 2.3.3.3 were taken into consideration. It should be noted here that although these lists were prepared for the English language, they make valuable implications for other languages as well. In the current study, these lists provided test writers with a guideline for choosing vocabulary and determining the appropriate level of difficulty for vocabulary in the spoken texts. Figure G1 shows the CEFR descriptors that refer to vocabulary knowledge (See Appendix G for Figure G1). According to the CEFR descriptors shown in Figure G1, in the current test, at lower levels, A1 and A2, mostly knowledge of words related to concrete needs and immediate needs are emphasized. In the corresponding tasks, words related to school, school environment, health, courses, course requirements, etc. were used in a simple and everyday language. Although sometimes topics can determine the kinds of words that may be used, the context and the discourse mode also influence the word choices. For example, the topic health can have a lot of sub-topics such as a simple headache or a complex disease, which requires low frequency and difficult words for discussion.

In the tasks in this study, the sub-topics and contexts did not require low frequency words. Instead, simple, everyday needs of students were taken into account while choosing words. However, the words “röntgen” or “ilaç” in A1 level text may be seen as a counter example to this situation. The word “ilaç” is the 1477th word with an observed frequency of 3380 prepared in the word frequency lists prepared as a part of the Turkish National Corpus (Aksan et al., 2012). This shows that this word and also the word “röntgen”, which is a low frequency word related with medicine science, are not used very frequently by language learners at lower-

levels of proficiency. Except for these words, the lexical coverage of the other items in A1 and A2 level task is mainly presumed to follow the requirements of the proficiency levels according to the CEFR specifications. At B1 and B2 levels, the range of vocabulary was considerably more diverse and at these levels listeners are supposed to have a sufficient repertoire of vocabulary to be able to cope with concrete and certain abstract topics as well as some idiomatic phrases and colloquial words. A more detailed investigation of the vocabulary used in the spoken texts is given in sections 4.3.3.4 and 4.3.3.5 under the discussion of the nature of information and content knowledge.

4.3.3.7 Grammatical resources

In the Reference Level Descriptors, possible grammatical structures that are likely to go with the functions and notions at different proficiency levels are suggested. The suggestions for the grammatical structures were considered while preparing the test tasks in this study. Moreover, at the earlier stages of the study, some course books used in TFL classes had been analyzed and common structures had been identified to be used in the texts. In addition to these, syllabi of Turkish as a foreign language courses at certain universities were examined to cross check the findings. After all these examinations, listening scripts were developed or adapted. Moreover, before the second pilot examination, expert opinion was also received from the instructors who are teaching TFL at Boğaziçi University. For example, one important feedback that was given was to delete –DIr suffixes at the end of the nominal predicates in the items in A1 task. This suffix is not taught at lower levels to students, so it might cause difficulty to test takers. An example would be the question in the A1 level task “Kaç tane öğrenci belgesi ücretsizdir?” was changed into “Kaç tane öğrenci belgesi

ücretsiz?”. Other small changes were also made in different tasks to better fit the grammatical properties of the text to the levels.

The overall complexity of the sentences in the current test is stated in the test specifications (See Appendix D for the test specifications). A gradation in the difficulty level of sentences is aimed across tasks. For A1 level only simple sentences, for A2 level mostly simple sentences, for B1 level a combination of simple and complex sentences with cohesive devices and linkers, and for B2 level many complex sentences were included. Below can be found some specific examples from each task to support our claims regarding the presumed grammatical difficulties of the texts.

The A1 text is generally made up of only simple, short sentences; however, three examples of adverbial clauses are observed in the text. The adverbial clauses are written in italic in the sentences below.

“Daha sonra başınız ağrırsa diye size ağrı kesici bir ilaç veriyorum.”

“Eğer kendinizi kötü hissederseniz, tekrar gelin lütfen.”

“Gerekirse röntgen çektirirsiniz.”

Among these, only one sentence includes an answer to an item. The sixth item in A1 level task targets the word “ilaç” but this word is not in the adverbial clause, but in the main clause. The other sentences are not targeted by any of the items. Normally, at this level the use of adverbial clauses are not seen often, but since this is a doctor-patient conversation, the use of conditional sentences, which are among the most frequently used adverbial clauses, were thought to be authentic and therefore included in the text. Still, the presence of these adverbial clauses seems to have increased the difficulty of the text as also observed in the results for the perceived difficulty of the texts by the test takers in Table 15 and the mean scores of

the total task scores in the second administration for the lower-level test takers in Table 24 in the investigation of research question 3. Therefore, these adverbial clauses need to be simplified or omitted from the dialogue. Another solution could be to choose a different topic which does not require adverbial clauses and low frequency words such as “ilaç” and “röntgen” for A1 texts.

In A2 level task, the structures are mostly simple again with the addition of some coordinate clauses such as “but”. Sometimes the sentences are long because they include small lists, but this does not affect the grammatical difficulty of the sentences. Thus, they are not expected to be too challenging for test takers at this proficiency level. Some examples of the coordinate clauses in A2 level task are italicized in the following sentences:

“Haftada üç saatmiş, hepsi peş peşe yapılacaktı *ama* yirmi dakikalık bir ara olacaktı.”

“Önce 206 nolu sınıf dedi *ama* sonra değiştirdi.”

“Aslında final tarihini de söyledi *ama* 6 Haziran mı 9 Haziran mı dedi, tam hatırlamıyorum.”

It can be seen here again that A2 level text includes simpler sentences than A1 level text, which contains adverbial clauses. Therefore, we can argue that the modifications suggested for A1 level text can improve the gradation across the tasks in terms of grammatical difficulty.

At B1 level, both simple and complex sentences were used together to increase grammatical difficulty. The complexity of the sentences was increased with coordinate clauses, and complex clauses such as nominal, adjectival and adverbial clauses. Below can be seen some extracts from the text which bear examples of coordinate and complex clauses, which are italicized.

“Biz öğlen gittik *ama* akşama kadar açık sanırım.” (coordinate clause)

“Aslında hiç yapasım yok *ama* yapmam lazım *yoksa* dersten kalacağım.” (coordinate clause)

“*Yeni üyelere hoş geldin partisi yapacaklarını* söylediler.” (nominal clause)

“*Cumartesi günü yorulduğum için* pazar günü bütün gün dinlendim.” (adverbial clause)

“*Eğer kızgın görünüyorsa* bir şey söyleme sakın” (adverbial clause)

“*Ocak ayında arkadaşlarla tatile gittiğimiz için* evde fazla kalamadım.” (adverbial clause)

“*Düşüne gitmeden önce* yetiştirdim *ama* çok hızlı yazdım.” (adverbial and coordinate clauses)

As can be seen from these examples, the grammatical complexity of the sentences is dramatically different from the A2 level text. Therefore, the frequency of using coordinate and complex clauses is supposed to comply with the target proficiency level.

At B2 level, the sentences get much longer with more occurrences of coordinate and complex sentences. At this level, learners are supposed to handle most of the grammatical structures of the language; therefore, the difficulty of the grammatical structures was not the primary concern. In addition, as Conrad (1985, in Elliott & Wilson, 2013) states, grammatical complexity does not affect comprehension of learners at higher levels as much as semantic load. Still, from a contextual validity point of view, the grammatical structures used in the texts should be representative of those learners are likely to encounter in TLU domain; thus, the texts were developed according to the requirements of the target level. There are

many coordinate and complex clauses and complex phrases at this level. Some examples are presented and italicized below:

“Beşiktaş Belediyesi ve Çocuk Masalları Akademisi’nin 4-5 Haziran tarihlerinde Akatlar Sanatçılar Parkı’nda düzenlediği Masal Şenliği’ni, etkinlik koordinatörü Ayşegül Dede’yle konuşacağız.” (adjectival clause)

“Masal Şenliği bizim 3 senedir üzerinde çalıştığımız bir proje.” (adjectival clause)

“Masallar çok etkili bir iletişim aracı ve bu sıralar masallardan esinlenen pek çok popüler sinema filmi ve televizyon dizisi görebiliyoruz.” (coordinate and adjectival clauses)

“Masalların ve masal anlatma geleneğinin korunup yaşatılması ve bunların gelecek kuşaklara aktarılmasına çok önem veriyoruz.” (nominal clause)

“Hem etkinliklerimiz sayesinde hem de katılımcılarımız sayesinde şenliğin bayağı gündeme oturacağını, çok konuşulacağını düşünüyoruz.” (nominal clause)

“Ankara’dan Somut Olmayan Kültürel Miras Müzesi gelecek yine bize eşlik etmek için.” (adverbial clause)

At B2 level due to its long sentences and complex grammatical structures, the text might sound unauthentic although it was slightly edited from a genuine interview. In order to be able to give the text a more authentic appearance, fillers such as “peki, yine, tabi ki, etc.” were used as well as some reversed sentences. The difficulty levels of the texts were also evaluated by the test takers on a four-point scale from 1 (*too easy*) to 4 (*too difficult*) in the third section of the task evaluation questionnaires. The results for the tasks are presented in Table 15. Based on the information in Table 15, it can be stated that there is variation and an expected difficulty cline across the tasks. The only exception to this variation is for A1 level task. This task is considered to be more difficult than A2 level task by the lower-

level test takers. As explained before, the reasons for this may be the relatively more difficult grammatical structures in A1 text. For the higher-level test takers, the gradation across A2 and B2 texts is smooth. Moreover, none of the texts were evaluated as too difficult or too easy although B1 level task was thought to be slightly difficult for the lower-level test takers. Still, it can be concluded that the tasks are generally considered close to moderate difficulty.

Table 15. Test Takers' Perceptions of Text Difficulty

Task	Mean Scores for lower-level test takers	Mean Scores for higher-level test takers
A1 Level Task	2.34	-
A2 Level Task	2.03	1.46
B1 Level Task	2.68	1.92
B2 Level Task	-	2.30

4.3.3.8 Functional resources

The functional dimensions of the texts in the current test are examined according to the categories of functional language in the Reference Level Descriptors as explained in section 2.3.3.3. Some examples from the recordings are shown in Table H1 in order to demonstrate the suitability of the demands of the tasks in terms of functional language for the different proficiency levels (See Appendix H for Table H1 which shows the functional dimensions of the listening texts). These examples from the listening texts in the tasks indicate that the variety of functional languages used in the tasks increases across different proficiency levels; therefore, the texts in this study are considered to contain a sufficient amount of various functional expressions.

4.3.4 Task demands (Interlocutor)

In addition to the linguistics demands of the tasks, interlocutor demands also need to be discussed as they hold great importance in terms of task demands. Next, aspects

of interlocutor demands will be explored with reference to the tasks in the current test.

4.3.4.1 Speech rate

In order to assess the speech rate of the recorded texts in the present study, the word per minute (wpm) and word per second (wps) values were calculated for each text and shown in Table 16.

Table 16. Speech Rates across Different Tasks

Tasks	Words per minute (wpm)	Words per second (wps)
A1 Level Task	126.6 words	2.11 words
A2 Level Task	110.4 words	1.84 words
B1 Level Task	134.4 words	2.24 words
B2 Level Task	113.6 words	1.89 words

It can be seen in Table 16 that there are variations across levels. It is normally expected that speech rate increases with the proficiency levels. However, in this test, A1 level task bears more words per minute and second than A2 and B2 level tasks. This indicates that the speech rate at A1 level is much higher than expected and therefore should be recorded again in line with the findings. It is also revealed that the total number of words in this task is more than that of A2 task. However, it should be noted here that as it was discussed in section 4.3.3.7 and it will be discussed in section 4.4.2.1 further in the study, these issues may have contributed to the difficulty level of A1 task and therefore should be controlled in the further use of the test.

Another issue that needs to be explained here is the relatively low number of words per minute in B2 level task. Normally at this level the speech rate is expected to be much faster with more words per minute and second. The situation in this task

may result from the content and linguistic load of the recorded text. Due to the grammatical complexity of the text with longer sentences, less frequent structures and words, and embedded clauses, etc. the speaker who recorded the B2 text may have chosen to read the text more slowly than expected in order to be able to naturally convey the complex and long messages more clearly. In this respect, it sounds logical because even native speakers tend to make small pauses between words and sentences when they are pointing to an important or complex piece of information in an interview as opposed to a social communication between people familiar to each other (acquaintances). In addition, when the length of the recorded text is taken into consideration, it is only normal that a speaker may not speak at the same pace for such a long time. Therefore, this situation is not considered as a major problem in the test because of the already demanding cognitive and linguistic load of the B2 level text.

In addition to the findings above, the test takers' evaluations of the texts were also analyzed in terms of speed. In the third section of the task questionnaires, the test takers evaluated the speed of the recordings by choosing one of the following: "1=slow", "2=normal" and "3=fast". The results of the questionnaire are demonstrated in Table 17.

Table 17. Mean Scores for the Speed of the Recordings

Task	Mean scores for lower-level test takers	Mean scores for higher-level test takers
A1 Level Task	2.40	-
A2 Level Task	2.09	1.46
B1 Level Task	2.87	2
B2 Level Task	-	2

According to the results in Table 17, A1 level text is considered to be faster than the A2 level text by the lower-level test takers, which is parallel to the results of the word count analysis mentioned earlier. This shows that A1 level text should be

recorded with a slower pace in the future versions of the test. B1 level text is considered to be the fastest recording by the lower-level group as expected. For the results for the higher group of test takers, A2 level text was regarded as slower than B1 and B2 level texts, which was the anticipated result. Another finding to be pointed out is the difference between the lower and higher groups of test takers. The average scores for both groups differ a lot, especially for B1 level text. The texts were thought to be considerably slower by the higher-level test takers when compared with the lower-level test takers' perceptions. This indicates that the texts can make a distinction between test takers at different levels of proficiency.

Buck (2001) argues that research results generally demonstrate that faster speeches are more difficult to comprehend. Since speech rate can impact on understanding the texts clearly and adequately, the test takers were also requested to state their opinions on the comprehensibility and audibility of the texts in the task evaluation questionnaires. The test takers evaluated their opinions about the following statements on a five-point scale, where "1" meant "definitely agree" and "5" meant "definitely disagree":

Statement 1: "The recording was comprehensible."

Statement 2: "The recording was audible."

The results of the test takers' perceptions of these two statements were given in Table 18 and Table 19 respectively.

Table 18. Test Takers' Perceptions of the Comprehensibility of the Texts

Task	Mean scores for lower-level test takers	Mean scores for higher-level test takers
A1 Level Task	1.62	-
A2 Level Task	1.31	1.23
B1 Level Task	1.21	1.30
B2 Level Task	-	1.07

Table 19. Test Takers' Perceptions of the Audibility of the Texts

Task	Mean scores for lower-level test takers	Mean scores for higher-level test takers
A1 Level Task	1.81	-
A2 Level Task	1.25	1.23
B1 Level Task	1.5	1.30
B2 Level Task	-	1.14

It can be seen from these results that there are variations across different tasks and some of the findings reflect the previous ones. For instance, A1 level text got higher mean scores in terms of comprehensibility and audibility from the lower-level test takers when compared to the other texts. This means that the lower-level test takers found A1 task slightly more challenging in terms of comprehensibility and audibility than they did the other tasks. This situation may result from the relatively higher number of words in A1 level text and the faster speech rate as discussed before. Therefore, although the result for A1 level task is actually on the positive side of the scale, a revision definitely needs to be made in the text at this level in order to create a balance across different proficiency levels. Another observation is that B2 level text has lower mean scores than the other texts and this can be explained by the lower wpm and wps values shown in Table 16 and the comparably clearer speech of the interlocutor that might have been adjusted due to the contextual and linguistic load of the text. However, despite some unexpected results, the results demonstrate in general that both groups of test takers mostly agree on the comprehensibility and audibility of the texts.

In conclusion, A1 level text needs to be modified to have a better gradation across tasks in terms of the speech rate, and the results for B2 level text is not considered to be a major problem due to the contextual and cognitive load the task imposes upon the test takers.

4.3.4.2 Variety of accent

Since this test is targeted for learners of TFL, only modern standard Turkish dialect used in Turkey is used for test purposes. Non-standard regional accents are not preferred since test takers are thought to be unfamiliar with them. Non-native speaker accents are not included in the test, either. However, in case of a context which requires a non-native speaker accent, it is possible to use it. For example, a conversation between two exchange students communicating in Turkish, or an exchange student conversing with a person around school environment such as a professor, a salesperson, a librarian, etc. would demand a non-native speaker accent. As a result, it should be noted that the TLU domain determines the requirement of accent. Especially when Turkish is taught as a foreign language more extensively, it is sensible to assume that language tests will also become more international with more non-native speakers included in the listening texts. However, whether the particular accent is generally available to the takers should be evaluated carefully.

4.3.4.3 Acquaintanceship

The texts in the study were recorded by people whom the test takers were not familiar with. However, the speakers were not considered to create problems of misunderstanding or incomprehensibility. The Turkish instructors also gave feedback regarding this issue and reported that the voices were sufficiently clear to facilitate comprehension.

4.3.4.4 Number of speakers

All of the texts in the current test are interactive and have two speakers. In the current test, a revision may be needed due to its lack of variety in terms of the

number of speakers. Some texts with one or multiple speakers and a lecture from a single speaker, especially at C1 level, may be added to create different contexts and discourse modes in the test. Moreover, the relationship between the speakers is generally stated in the instructions for each task as well as the basic context, which is likely to shape the formality of the language and the way the conversation will continue.

4.3.4.5 Gender

The genders of the speakers in the texts are mixed and balanced with no specific cultural or socio-economic background. To avoid cultural bias, a mixture of both genders is included in the test. From a testing perspective it also makes it easier for test takers to distinguish the voices when speakers are of opposite genders, especially in interactive texts.

4.3.5 Conclusion for the investigation of context validity

The second research question aimed to explore issues related to the contextual features of the tasks designed as part of the present study. The task setting, administrative setting, linguistic demands and interlocutor demands of the tasks have been scrutinized both theoretically and statistically. Suggestions for revision were proposed when the tasks did not meet the expected requirements. Overall, it can be stated that the criteria for context validity have been taken into account while preparing the tasks and results supporting this have been gathered as a result of two pilot sessions. Through the argumentation and descriptive data, this part, as an answer to the second research question, have been able to establish context validity evidence for the tests being developed for TFL learners. Next, in order to support our

validity claims about this test further, statistical analyses on tests and items will be carried out to investigate the third research question.

4.4 Investigation of scoring validity

Up to this point, cognitive and contextual validity evidence that was collected before and after the administration has been presented. In this section of the chapter, a posteriori evidence for our scoring validity claims about this test is presented. To this end, the statistical analyses conducted on the test takers' performances in the first and the second piloting sessions are investigated in order to explore how well the items and the tasks functioned in the test. The third research question and its sub-questions, which aim to explore scoring validity issues, are shown below:

Research Question 3: How well do the test and the items function in terms of scoring validity?

- a. Do the values for central tendency measures of the tasks and item analyses based on the test takers' performances support that the test is functioning well?
- b. Does the test measure the listening ability of learners of TFL reliably?

As explained thoroughly in Chapter 3, classical item analysis procedures, i.e. measures of central tendency, reliability and item analysis, are applied in this chapter to analyze the test scores obtained both in the first and second administration of the test. For the first piloting, the results of the classical item analysis for the whole test are demonstrated and the reasons for the relevant changes made in the tasks for the second administration are discussed. For the second piloting, the results of the classical item analysis for the test are examined for two different groups of test takers: A1-B1 level tasks for the lower-level students and A2-B2 tasks for the higher-

level students. In addition to the classical item analysis, if possible problems with the items and the tasks have been observed, the reasons and certain suggestions to mitigate them are provided. These are going to be discussed under research question 3 through issues detailed in sub-questions 3a and 3b in the following pages.

4.4.1 Item statistics from the first pilot administration

Firstly, the test and item statistics from the first pilot administration are discussed to lay the basis for the changes made for the second pilot examination. The descriptive statistics and the score distribution data for the first pilot administration of the whole test can be seen in Table 20. The table demonstrates that the mean for the overall test scores is 19.31 out of 42 (45.9%). There is no cut-off score determined for this test's results; however, still it can be argued that the mean score is slightly lower than expected since it is below 50%. In terms of the skewness and peakness of the scores, the values for skewness (0.106) and kurtosis (-1.258) are within the acceptable range. The calculated alpha value for the test scores (.954) indicates a very high level of internal consistency. Therefore, based on these findings and Figure 6, which shows the distribution of the scores in the first administration, it can be concluded that the test scores for the first pilot administration are distributed fairly normally with low skewness and relatively higher flatness values and the test has strong claims for reliability. The statistics for the overall test scores can be considered satisfactory.

Table 20. Descriptive Statistics of the Total Test Scores from the First Pilot Administration

N	Item n.	Range	Min.	Max.	Mean	SE	Std.	Skewness	Kurtosis	Alpha
55	42	39	0	39	19.31 (45.9%)	1.57	11.70	0.106	-1.258	.954

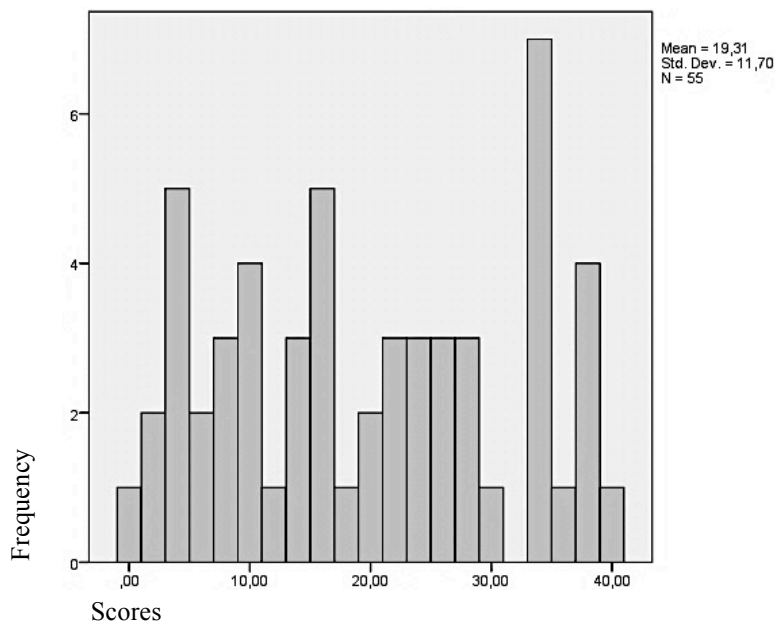


Figure 6. Distribution of the test scores in the first pilot administration

The tasks are also analyzed in terms of their mean scores. This test is composed of five tasks each of which has a different level of proficiency. Therefore, we also need to investigate if these tasks are ordered according to their expected difficulty levels. It is hypothesized that A1 level task will have the highest mean score whereas the most difficult task, C1 level task, will have the lowest mean score in the test. Table 21 demonstrates the mean scores for the total scores on each task and it can also be observed that the mean scores do not reflect the expected order for the proficiency levels. The tasks in the test are ordered in the following way from the easiest to the most difficult: B1, A2, A1, C1 and B2.

Table 21. Mean Scores for the Tasks in the First Pilot Administration

Task	Mean scores	Mean scores out of 100
A1 level task	2.96 /6	49.33
A2 level task	3.31 /6	55.1
B1 level task	5.96 /10	59.6
B2 level task	3.33 /10	33.3
C1 level task	3.75 /10	37.5

Individual item analyses are also carried out to have a better idea of how the items work together. Table 22 shows the results of the item analysis for the first pilot study. The item mean (IF), item discrimination (ID, Corrected-Item Total Correlation-CITC) and alpha if item deleted (AIID) values need to be considered to demonstrate the extent to which the items function well in the test. According to the ID and AIID values of the first administration shown in Table 22, one item seems to create problems in the overall test. The second item in B1 level test (B1I2) has a low value for item discrimination (.223), which is below the acceptable value, .30. In addition, when this item is deleted from the test, the overall reliability of the test increases (AIID=.955). This item also has the highest mean score among all the other items and this situation may have led to low discrimination and reliability values for this item. In addition, the third and fifth items in C1 level task have too low IF values, (.13) and (.11) respectively. Therefore, these items were omitted in the second version of the test because of their low item values, the modifications made on B1 level task after the first pilot administration due to the concerns mentioned in sections 4.2.3.1 and 4.2.3.2, the exclusion of C1 level task from the test as discussed in section 4.2.5 and some other problems mentioned below. Apart from these items, the other items in the test seem to contribute to the test in terms of discrimination and reliability.

The findings presented for the first pilot administration signifies crucial issues in terms of task difficulty and also the target cognitive sub-skills aimed at in the tasks. The cognitive concerns for these tasks are discussed in section 4.2 and it is stated that although the reliability and discrimination values of the items and the tasks were quite satisfactory after the first administration, due to problems with the target proficiency levels and the cognitive processes, the tasks were modified greatly.

Tasks which were considered to be easier than expected were altered to make them more difficult and appropriate for the proficiency levels. Similarly, the tasks which were found to be too difficult for the target proficiency levels were simplified in terms of items, cognitive sub-skills and language. Therefore, we can maintain that the analysis of the mean scores provided valuable findings and implications for modifications towards the second version of the test.

Table 22. Item Analysis Statistics in the First Pilot Administration

Items	IF	CITC	AIID
A1I1	.55	.463	.953
A1I2	.60	.540	.953
A1I3	.44	.631	.952
A1I4	.62	.500	.953
A1I5	.36	.755	.952
A1I6	.40	.568	.953
A2I1	.58	.546	.953
A2I2	.65	.493	.953
A2I3	.49	.702	.952
A2I4	.62	.653	.952
A2I5	.47	.581	.953
A2I6	.49	.709	.952
B1I1	.47	.641	.952
B1I2	.78	.223	.955
B1I3	.69	.364	.954
B1I4	.22	.534	.953
B1I5	.62	.598	.953
B1I6	.75	.473	.953
B1I7	.75	.380	.954
B1I8	.62	.704	.952
B1I9	.60	.664	.952
B1I10	.47	.677	.952
B2I1	.49	.481	.953
B2I2	.36	.510	.953
B2I3	.27	.527	.953
B2I4	.33	.534	.953
B2I5	.24	.356	.954
B2I6	.27	.381	.954
B2I7	.53	.621	.952
B2I8	.38	.625	.952
B2I9	.20	.385	.954
B2I10	.25	.407	.954
C1I1	.64	.662	.952
C1I2	.36	.709	.952
C1I3	.13	.529	.953
C1I4	.55	.609	.953
C1I5	.11	.406	.954
C1I6	.47	.575	.953
C1I7	.42	.582	.953
C1I8	.29	.750	.952
C1I9	.40	.720	.952
C1I10	.38	.666	.952

In order to have a more detailed idea of the order of the tasks, the individual mean scores for the items were also ordered. Table I2 shows the order of the items in the first pilot administration from the easiest to the most difficult (See Appendix I for Table I2). According to the order shown in Table I2, it can be seen that the items of B1 level task are at the top of the list of mean scores or very close to the top, which also supports that it is the easiest task in the test with the highest mean score. The items in B1 level task are followed by items from different tasks but mostly from A2 level and A1 level tasks. The items in B2 and C1 level tasks seem to be more difficult than the others, but the expected order is not seen here, either. B2 level task is proved to be more difficult than C1 level task as there are more items from B2 level task at the end of the list than C1 level task items. Thus, it can be stated that the order in Table I2 shows parallel results to the order of the mean scores for the total task scores in Table 21.

In summary, it may be impossible to expect a perfect order in a test in terms of item mean scores. Task difficulty is closely related with text and item sophistication which are in turn decisive in cognitive operations to be used in response to tasks. Task difficulty is, therefore, a major concern in theory-based and context validity. No matter how good the scoring validity of a test is (reliability), it is impossible to disregard unsuccessful operationalizations of language skills in a test and thus the test had to be revised. Therefore, some changes were made regarding certain items and texts. Due to these results and cognitive validity concerns as explained in section 4.2, A1 level task was kept, but the dialogues and items were revised considerably to make the task easier and more appropriate for the level. A2 level task became B1 level task in the new version of the test, as it was more difficult than B1 level task in the first administration. Accordingly, B1 level task became the

new A2 level task in the second pilot administration. B2 and C1 level tasks were completely omitted from the test and only one new task at B2 level was developed from scratch. Next, the statistical analysis results for the second administration are presented in order to show the results of these changes.

4.4.2 Item statistics from the second pilot administration

As mentioned before in Chapter 3, the tasks in the second administration were given to two different groups of test takers due to technical and practical reasons imposed by the TFL course instructors. TFL teachers suggested that A1 level task should not be given to the higher-level test takers, and B2 level task should not be given to the lower-level test takers. Therefore, A1, A2 and B1 level tasks were delivered to one group of test takers named as lower-level test takers (participants in Classes 20, 21 and 25 in the Turkish Language and Culture Program) whereas A2, B1 and B2 level tasks were given to another group of test takers named as higher-level test takers (participants in Classes 30 and 31 in the Turkish Language and Culture Program). The results are discussed separately for these two different groups of participants.

4.4.2.1 Statistical analysis results for the lower-level test takers

The descriptive statistics for the test scores of the lower-level test takers are given in Table 23 and the distribution of the scores are shown in Figure 7. The distribution of the scores do not differ much from the first administration in terms of skewness and peakness. The total scores are slightly more positively skewed in the second administration with a skewness value of .401 and the kurtosis value (-1.148) seems close to that of the first administration (-1.258), which means the scores do not show peakness but flatness. One big difference is related with the reliability coefficient.

The alpha score in the second administration (.729) is considerably lower than the alpha score for the first pilot study (.954). However, .729 as an alpha value is still within the accepted range and indicates a good reliability score. The reason for the decrease in the alpha value can be attributed to the smaller sample size, the homogeneity of the participants and the lower number of items and tasks in the second administration and therefore, a further study with more participants can yield more dependable results in this respect. Overall, when the data in Table 23 and Figure 7 are considered, the distribution of the test scores in the second administration can be considered to reflect a close-to-normal distribution with some skewness and flatness.

Table 23. Descriptive Statistics of the Test Scores in the Second Pilot Administration for the Lower-level Test Takers

N	Item n.	Range	Min.	Max.	Mean	SE	Std.	Skewness	Kurtosis	Alpha
16	20	12	6	18	11.44 (57.2%)	.978	3.91	.401	-1.148	.783

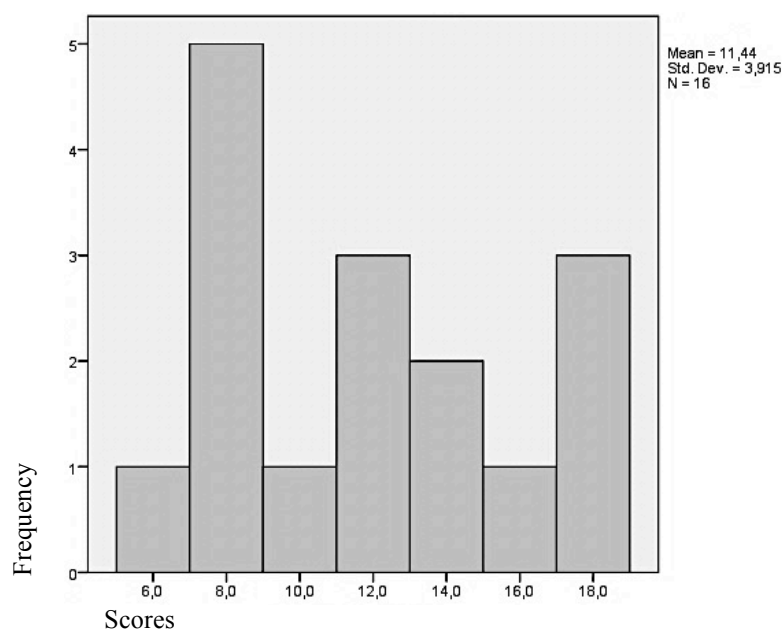


Figure 7. Distribution of the test scores in the second administration for the lower-level test takers

We also need to investigate whether the tasks are ordered according to their expected proficiency levels. Table 24 shows the mean scores of each task for lower-level test takers. The results in Table 24 show that A1 level task is much more difficult than A2 level task, which is contrary to our expectations. As mentioned before, A1 level task was found to be more difficult than A2 level task in terms of text length, wpm calculations, grammatical difficulty, and test taker's evaluation of text length and text difficulty. Based on these findings and the major difference between these two tasks in terms of mean scores, we can argue that A1 level task needs to be revised. For this revision, item analysis statistics that are discussed below needs to be considered and the suggestions made for text length and grammatical difficulty should be taken into account. B1 level task is the most difficult task for this group of test takers and therefore, does not seem to require any modifications in terms of difficulty.

Table 24. Mean Scores for the Tasks in the Second Pilot Administration for the Lower-level Test Takers

Task	Mean scores	Mean scores out of 100
A1 level task	3.13/6	52.16
A2 level task	5.25/8	65.62
B1 level task	2.88/6	48

In addition to the statistical analyses of the items, in the second pilot administration, qualitative data from the test takers were also collected. The test takers completed task evaluation questionnaires for each task and evaluated the difficulty levels of the items in each task on a scale ranging from 1 (*too easy*) to 4 (*too difficult*) in the second section of the questionnaires. The mean scores for these evaluations are calculated separately for the lower-level and higher-level test takers. These mean scores are demonstrated in Table 25. The results of the questionnaires

are parallel to those obtained from the statistical analysis. The test takers perceived A2 level task as the easiest and B1 level task as the most difficult.

Table 25. Test Takers' Perceptions of Task Difficulty in the Second Administration

Task	Lower-level test takers	Higher-level test takers
A1 level task	2.20	-
A2 level task	1.95	1.52
B1 level task	2.72	2.11
B2 level task	-	2.70

The fact that A1 level task was considered more difficult than A2 may be caused by a few other factors in addition to the factors mentioned before such as text length, speed rate and grammatical difficulty. Firstly, the number of speakers in A1 level task was higher since there were three short conversations in the task. The test takers may have had difficulty adapting to the voices of different speakers. Moreover, listening to three short conversations may have created some problems for the test takers since they needed to create a context in their minds for each of the short dialogues. Therefore, for future versions of this test only one relatively longer text can be played and the amount of redundancy and repetitions can be increased to make the test more level appropriate. In addition, the test takers may have deemed the linguistic and cognitive difficulty level of A2 level task easier. In order to mitigate this problem, A2 level listening text can be modified in order to convert it into a comparably more challenging text. This can be done via slightly increasing the difficulty level of the vocabulary and grammatical structures, and the target cognitive skills can be modified as discussed in sections 4.2.2.1 and 4.2.2.2. More items which target sentence-level factual information can be incorporated in the task to distinguish it from A1 level task. These changes are likely to create a balance between the tasks in terms of difficulty. However, the IF, ID and AIID values also need to be taken into account while making such revisions.

After analyzing the total scores for the tasks and the test, the next step is to analyze the items in terms of item facility, discrimination and reliability. Table 26 shows IF, ID and AIID values for the items in the second administration for the lower-level group. When these values are analyzed, it can be seen that items A1I1, A2I1, A2I2, A2I4, B1I2, and B1I5 have very low discrimination values (-.113, .000, .000, -.380, .183 and -.073 respectively). In addition, the same items have a negative effect on the overall reliability of the test because when they are deleted from the test, the alpha score for the test increases.

The low values for item discrimination and reliability may stem from two major reasons in the overall test: One reason could be the considerably lower sample size in this administration. In the first version of the test, 55 students took part in the study whereas two different groups of test takers participated in this version. The number for the lower-level test takers is 16 and therefore, the smaller number of test takers may have affected the statistical analysis results. Another reason could be that the test was given to a homogenous group of test takers and there was little variability in terms of ability among the test takers. The lower-level group was composed of Turkish learners at or below intermediate proficiency level; therefore, items may not have discriminated well enough between these test takers with similar proficiency levels. This indicates that the item statistics should be considered very reliable and should make little impact on the test developer's decisions. In addition to the general reasons for low ID and AIID values, we also need to scrutinize the individual items in terms of, text characteristics, item characteristics and test takers' responses in order to have a clearer idea of the source of the problems. Therefore, each of the items mentioned above is examined below with reference to the possible sources of low ID and AIID values.

Table 26. Item Analysis Statistics for the Lower-Level Test Takers in the Second Pilot Administration

Items	IF	CITC	AIID
A1I1	.81	-.113	.799
A1I2	.13	.567	.763
A1I3	.69	.252	.780
A1I4	.81	.414	.770
A1I5	.25	.360	.773
A1I6	.44	.713	.745
A2I1	1.00	.000	.785
A2I2	1.00	.000	.785
A2I3	.56	.612	.753
A2I4	.94	-.380	.801
A2I5	.38	.505	.762
A2I6	.31	.310	.776
A2I7	.75	.237	.781
A2I8	.31	.509	.762
B1I1	.38	.584	.756
B1I2	.44	.183	.786
B1I3	.56	.495	.763
B1I4	.56	.419	.768
B1I5	.56	-.073	.803
B1I6	.38	.790	.740

Firstly, A1I1 is investigated in terms of its text characteristics, item characteristics and test takers' responses. The A1 text is assumed to be generally simple and appropriate for A1 level except for some adverbial clauses as explained before in sections 4.3.3.6 and 4.3.3.7. When the texts are examined, it can be seen that there are no distractors for A1I1 and since the text is not challenging, almost all test takers except for three could answer the item correctly. The three test takers who could not provide the right answer had moderate and high total scores in the test (11 and 13 out of 20) whereas some test takers who answered the items correctly got very low scores (6, 7 and 8 out of 20) in the test. This shows that the item could not discriminate well between high and low achievers, which in turn impacted the ID and AIID values of the item. The item values can be improved by adding some more distractors in the text so that those with higher language abilities will get the answers correct.

Secondly, A2I1 does not discriminate at all between the test takers as everyone could answer this item correctly. It has an IF value of 1.00 and lowers the reliability of the test (AIID=.785). When we look at the A2 text, we can see that there are no distractors for the correct answer in the text. This item is a multiple-choice question with three options and it can be seen that the incorrect options are not heard in the listening. The only word that can be heard as a distractor from the options in the text is “ödev”; however, this word is not uttered by the person who is supposed to tell the correct answer of the item. In order to improve the quality of this item, we need to rewrite certain parts of the text, especially the beginning, with better distractors and integrate these distractors into the speech of the person who gives the correct answer. The statistical analysis results for this item indicate another important finding. In section 4.2.2.1, it was argued that this item required inferencing skills; therefore, it was expected to cause difficulty to the test takers. However, due to its text and item characteristics, it did not yield the expected results. On the contrary, due to the lack of distractors, it was found to be one of the easiest items in the test. This again shows that difficulty is a combined result of cognitive and contextual features of the text, the task and response characteristics.

Similar to A2I1, A2I2 also has an ID value of .000 and IF value of 1.00, and it causes a decrease in the overall reliability of the test (AIID=.785). The reason for these low values seems to be parallel to those explained for A2I1. For A2I2, the test takers needed to answer a short-answer question which asked the day of the course. When the text is analyzed, it can be found that there are not any other day names, or words, that can distract the test takers. Only one day is mentioned in the text and it is the correct answer. Inclusion of distractors in the text for this item can increase the item quality, and IF, ID and AIID values.

A2I4 similarly has low item statistics due to the lack of distractors in the text. The question asks the surname of the author of the book and it is repeated in the text twice. Therefore, the absence of distractors and the redundancy in the text helped the test takers find the correct item. Furthermore, the only test taker who could not find the right answer got a high total score (16 out of 20) in the test, which may have lowered the ID score even more. However, the answer that this test taker gave was “Ahsan”, which is very similar to the correct answer “Aslan”, which indicates that the test taker was not distracted by any other words in the text. Therefore, the quality of this item can again be improved by adding distractors in the text.

Another item, B1I2 has low ID value (.183) and decreases the reliability of the test. It is a multiple-choice item with four options; therefore, distractor analysis of the options in this item is also analyzed along with the text characteristics. When the options and the text are examined, it can be seen that all of the options are integrated into the text; therefore, the listeners had to listen for all of the options. Distractor analysis shows that the numbers of lower-level test takers that chose the options A, B, C and D are one, two, six and seven respectively. This means that option C was preferred almost as much as option D, which was the correct answer. For the higher-level test takers², options A, B, C and D was chosen by one, three, two and nine test takers. Only the correct answer seems to attract most of the answers and none of the distractors are strong. In order to understand the reason for this, we need to examine the options. All of the options contain names of student clubs at a university and the test takers had to choose the student club whose activity the speaker attended on Saturday evening. The student clubs in the options are all mentioned in the text as a

² As it is explained in section 4.4.2.2, the results for the higher-level group will not be discussed separately due to the small sample size and the low item values because of the sample size. The discussion of the problematic items both for higher and lower-level groups is, therefore, given together.

part of the speaker's weekend activities. Therefore, one crucial aspect of the text was the time words mentioned in the text since they would give the listeners clues as to the time of the student club activities. While developing the text and the item, words indicating time such as "have breakfast", "around 12 o'clock", "in the afternoon" and "in the evening" were added in the text. However, after analyzing the text again, it was realized that the time for the activity of the club in option C (the strongest distractor for the lower-level group) was given *after* talking about the event. This may have confused especially the lower-level test takers since they did not hear the time of the event before and therefore, they may have considered that it took place on Saturday evening and chosen it as the correct answer. One solution to this would be to specify the time for the activity of the club in option C before mentioning it so that the listeners can follow the time sequence given in the text more easily. Another problem related with this item is the test takers' responses and their total scores. Most of the lower-level test takers who answered the item correctly (four out of seven) got lower total scores from the test (8, 12 and 13 out of 20) when compared with others and one test taker who got a very high total score (18 out of 20) answered the item incorrectly. These may also have resulted in low ID and AIID values. For the higher-level learners, it can be seen that four of the higher-level test takers who answered the item incorrectly got very high scores from the test (17 and 18 out of 22). Overall, the ID values seems to have lowered due to the number of test takers with high total scores who answered the item incorrectly and the number of test takers with low total scores who answered the item correctly. In order to mitigate this problem, we need to modify certain parts of the text related to the distractors of this item and make the time sequence in the text clearer.

Finally, analysis of item B1I5, a multiple-choice question with four options, demonstrates that the distractors in the item do not function well similar to item B1I2. Options A, B, C and D attracted one, three, nine and two answers from the lower-level test takers, and options A, B, C and D attracted 0, 0, 12 and 2 answers from the higher-level test takers respectively (Option C is the correct answer). When the text and the options are examined, it can be seen that all of the options are implied in the text; however, two of the options (A and D) were not uttered by the person who gave the correct answer. This may have helped the test takers eliminate these options. If these options are also integrated into the speaker's speech who gives the correct answer, the quality of the distractors can be enhanced. The total scores of the test takers are also analyzed. Six of the nine lower-level test takers who answered the item correctly got low total scores from the test (7, 8, 10, 11, 12 and 13 out of 20) and two of the seven lower-level test takers who answered the item incorrectly got very high total scores (17 out of 20). For the higher-level test takers, however, such a case is not observable; the test takers who gave the correct answer got high scores except for one test taker who got 13 of 22 and the test takers who answered the item incorrectly got low scores from the test (11 and 14 out of 22). This shows that the item could not discriminate very well between strong and poor learners in the lower-level group. As mentioned above, by revising the places of the distractors in the text, we can obtain better item statistics.

The items examined above need modifications due to their low ID and AIID values; however, there is one more item in the test which requires examination because of its too low IF value. A1I2 has the lowest mean score (.13) for the lower-level test takers. When the text is analyzed, it can be understood that the correct answer (Yedi lira 75 kuruş) to A1I2 was mentioned only once in the text. In addition,

the fact that the correct answer is consisted of two pieces of information, i.e. “Yedi lira” and “75 kuruş”, may have made it too challenging for the test takers. The reflection of this can be seen in the test takers’ responses. Eight test takers out of 16 did not provide any answers to this item and four test takers could only provide “75 kuruş” as the answer. This shows that the answer to this item needs to be made shorter and less complicated for the test takers and some repetitions can be added to make the item more level appropriate.

In addition to the individual item statistics, the ordering of the individual items should be examined to see if the items are ordered according to their proficiency levels. The ordering of the items according to their mean scores for the lower-level test takers can be seen in Table I3 (See Appendix I for Table I3). In Table I3, it is clear that the items in A2 level task are easier than those in A1 level task since most of the items in A2 level task are ordered at a higher position than the items in A1 level task. This also supports the findings mentioned earlier in various sections that A1 level task is more difficult than A2 level task. It is also surprising to see that items A1I2, A1I5, A2I6 and A2I8 have even lower mean scores than the items in B1 level task. However, the items except for A1I2 are within the acceptable ranges and they have acceptable ID and AIID values, modifications in these items are not our primary concern. Despite the irregularities in the order of the items, it is noteworthy that the items are better ordered in the second version of the test compared to the first version.

The suggestions put forward in this section are hoped to improve the descriptive scores for the whole test and the item statistics, and contribute to the reliability of the test and the items. In this way, we can observe gradation across the tasks and the results of the test can be interpreted more reliably.

4.4.2.2 Statistical analysis results for the higher-level test takers

The second group of test takers was the higher-level test takers. The descriptive statistic results for the total test scores of this group are shown in Table 27. In Table 27, it can be seen that the total scores are negatively skewed and more peaked, which indicates the higher number of people who got higher scores in the test. However, the values for skewness and kurtosis are close to “0” and can be considered as reasonable. The histogram shown in Figure 8 also demonstrates the peakness and skewness of the scores; however, based on the shape of the histogram we can discuss that the distribution of the scores are close to normal. The alpha score of the test is .642 for higher-level groups, which is lower than that for the lower-level group and the first administration. Yet, the reliability value can still be considered moderate.

Table 27. Descriptive Statistics of the Test Scores in the Second Pilot Administration for the Higher-level Test Takers

N	Item n.	Range	Min.	Max.	Mean	SE	Std.	Skewness	Kurtosis	Alpha
14	22	11	11	22	17.14 (77.9%)	.776	2.9	-.604	.487	.642

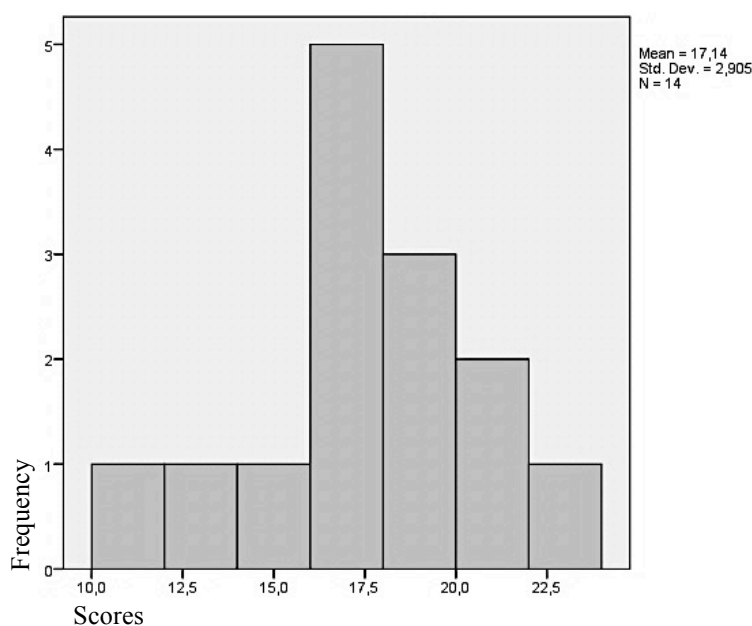


Figure 8. Distribution of the test scores in the second administration for the higher-level test takers

The order of mean scores of the tasks should also be demonstrated to provide an overall idea of task difficulty in the test. The mean scores for the tasks for the higher-level group are presented in Table 28. The findings show that the tasks are ordered in the predicted manner from the lowest proficiency level to the highest.

Table 28. Mean Scores for the Tasks in the Second Pilot Administration for Higher-level Test Takers

Task	Mean scores	Mean scores out of 100
A2 level task	7.50/8	93.75
B1 level task	4.71/6	78.5
B2 level task	4.93/8	61.62

The test takers' perceived task difficulty results from the task evaluation questionnaires are also shown in Table 25 with the results for the lower-level group in section 4.4.2.1. This also shows clearly that the tasks in the second administration were considered to be as difficult as their predicted proficiency levels for the test takers. That means that the tasks are ordered according to their proficiency levels as in Table 28. There is also a substantial difference between the perceived difficulty levels of the lower and higher groups of test takers for the same tasks. All of the tasks were evaluated as easier by the stronger test takers compared to the weaker test takers and this also provides important contextual validity to support the claims that A2 and B1 level tasks should function differently for weak and strong learners.

The item analysis statistics are also demonstrated to see the IF, ID and AIID values for each item in Table 29. However, these results will not be discussed in this chapter, as they are beyond normal due to the small size of the sample that was consisted of a very homogeneous group of test takers (n=14). Instead of discussing the unsatisfying results for the higher-level test takers, a set of hypothetical data will be presented. As mentioned earlier, the TFL instructors argued that the higher-level test takers could get the items in the A1 level task correctly and the lower-level test

takers could get the items in the B2 level task incorrectly; therefore, these two tasks were not administered to all of the participants. Based on the opinions of the TFL instructors, a hypothetical set of data where the higher-level test takers all got “1” for the items in A1 level task and the lower-level test takers got “0” for the items in B2 level task was formed and analyzed. In this way, we will be able to report findings that come from a bigger sample size. The results of the item analysis are shown in Table 30.

Table 29. Item Analysis Statistics for the Higher-Level Test Takers in the Second Pilot Administration

Items	IF	CITC	AIID
A2I1	1.00	.000	.644
A2I2	1.00	.000	.644
A2I3	.93	.330	.624
A2I4	1.00	.000	.644
A2I5	.93	-.078	.655
A2I6	.79	-.179	.678
A2I7	.93	.330	.624
A2I8	.93	.330	.624
B1I1	.57	.545	.582
B1I2	.64	.138	.644
B1I3	.93	.546	.607
B1I4	.93	.546	.607
B1I5	.86	.600	.590
B1I6	.79	.781	.556
B2I1	.64	.437	.600
B2I2	.29	.091	.649
B2I3	.71	.100	.648
B2I4	.50	-.077	.676
B2I5	.64	.375	.610
B2I6	.86	.042	.650
B2I7	.64	.027	.660
B2I8	.64	.196	.636

One of the first things that can be observed is the reliability coefficient that significantly increased to .924. The increase in the participants and the number of items significantly affected the alpha score. In contrast, the problematic items with low ID and AIID values still seem to be the same as the ones discussed for the lower-level test takers. The low values for these items except for B1I5 can also be observed for the higher-level test takers in Table 29. However, it should be noted that the ID

and AIID values for these items in Table 30 for the hypothetical analysis are not as low as for values for the lower-level test takers in Table 28. Furthermore, most of the items with low ID and AIID values in Table 29 are not seen in Table 30. Even though the values in Table 30 are not obtained as a result of the analysis of real test scores, they can still be considered to reflect, to some degree, the results that would have been received if the test had been administered to all the participants. The TFL instructors' evaluations of the test takers and the differences between the mean scores of the tasks for the different groups of test takers can be argued to support this consideration.

Table 30. Item Analysis Statistics for the Hypothetical Test Scores from the Second Administration

Items	IF	CITC	AIID	Alpha
A1I1	.90	.226	.925	.924
A1I2	.53	.866	.916	
A1I3	.83	.456	.923	
A1I4	.90	.427	.923	
A1I5	.60	.754	.918	
A1I6	.70	.752	.918	
A2I1	1.00	.000	.926	
A2I2	1.00	.000	.926	
A2I3	.73	.593	.921	
A2I4	.97	-.005	.926	
A2I5	.63	.637	.920	
A2I6	.53	.465	.923	
A2I7	.83	.332	.924	
A2I8	.60	.721	.919	
B1I1	.47	.429	.924	
B1I2	.53	.258	.927	
B1I3	.73	.581	.921	
B1I4	.73	.557	.921	
B1I5	.70	.338	.925	
B1I6	.57	.698	.919	
B2I1	.30	.689	.919	
B2I2	.13	.387	.924	
B2I3	.33	.668	.920	
B2I4	.23	.490	.922	
B2I5	.30	.677	.920	
B2I6	.40	.752	.918	
B2I7	.30	.607	.921	
B2I8	.30	.642	.920	

The order of the mean scores of the items for higher-level test takers also needs to be presented to detect whether any items are problematically ordered. This order can be seen in Table I4 (See Appendix I for Table I4). The results for the higher group of test takers can be considered more satisfactory when compared to the lower-level group and the first group of test takers. The items are much better ordered with only some items being more difficult than expected (items A2I6, B1I2, B1I1) and a few items being easier than expected (items B1I3 and B2I6). The rest of the list seems very straightforward and clearly shows the easiest and the most difficult items. Thus, the list can be said to reflect the statistical results for the total mean scores and also the task evaluations of the test takers.

4.4.3 Conclusion for the investigation of scoring validity

In the light of the discussions provided for the investigation of research question 3, it can be concluded that the statistical analysis results of the second administration provided less satisfactory results in terms of reliability and item discrimination despite the cognitive and contextual changes made in the second version of the test. This can be explained with the smaller sample size and homogeneity of the test takers in the second administration. Suggestions made for modification of items A1I1, A1I2, A2I1, A2I2, A2I4, B1I2, and B1I5 are expected to increase the IF, ID and AIID values of these items. On the other hand, the mean scores across tasks seem to have improved and the expected proficiency and difficulty levels of the tasks are reflected much better in the second administration. These results can be seen as a reflection of the modifications made in the tasks in terms of cognitive and contextual validity. Finally, the statistical analysis of the hypothetical data implies promising

results for the scoring validity of this test provided that it is administered to a larger group of test takers with heterogeneous language backgrounds.

4.5 Conclusion for Chapter 4

In this chapter, we attempted to provide evidence for the cognitive validity, context validity and scoring validity of our claims about the results of the present test. The test, test tasks and test specifications were analyzed theoretically and statistically. The theoretical discussions and empirical data provided us with crucial findings about the study and ways to improve the test were suggested accordingly. In the next chapter, the findings, the discussions and the suggestions mentioned in this chapter are summarized and limitations of this study as well as more suggestions for future research are presented.

CHAPTER 5

CONCLUSION

5.1 Summary of the findings

This test is developed to assess listening proficiency of learners of Turkish as a foreign language (TFL) and help determine the proficiency levels of the foreign students who learn Turkish at Boğaziçi University and the validation of it is conducted according to Weir's (2005) framework and Field's (2013) listening model. Three main research questions are investigated via theoretical and statistical analyses and a summary of the findings for these research questions are presented below.

5.1.1 Summary of the findings for the first research question

The qualitative and quantitative analyses of the cognitive requirements of the tasks demonstrated that A2 level task failed the cognitive requirements of the target proficiency level in terms of construct representativeness and comprehensiveness. Therefore, this task needs to be modified for the future versions of the test. Furthermore, the analysis for the tasks in the first administration (B1, B2 and C1 level tasks) showed that the difficulty level of the texts does not necessarily guarantee task difficulty or help trigger higher-level listening processes. Another important observation was that marking sentences as True/False/Not Given as task type in listening tests was not a very suitable means of assessing the listening skill due to the online nature of listening. Other than these observations, the tasks in the second administration generally seem to satisfy the cognitive requirements of the theoretical frameworks utilized in this study and a variation and gradation across tasks in terms of the listening processes they elicit can generally be observed. The

listening construct is assessed sufficiently with a variety of target cognitive processes and the construct validity of the test is enhanced when compared to the findings for the first administration of the test.

In addition to the task-specific findings mentioned above, the investigation of the first research question also demonstrated that task evaluation questionnaires provided invaluable data regarding the cognitive processes and listening sub-skills employed during listening and they showed the possibility of identifying these processes and sub-skills. Moreover, according to the task evaluation questionnaires, the lower-level test takers tended to use a more diverse range of sub-skills while responding to the items. These test takers may have attempted to compensate for their lack of comprehension by mixing up top-down and bottom-up processes. It might be much easier for higher-level test takers to answer items that target lower-level listening sub-skills and specific information; however, when lower-level test takers do not have the necessary linguistic knowledge or automatized lower-level listening sub-skills, they may employ other sub-skills to enhance their understanding.

5.1.2 Summary of the findings for the second research question

One essential conclusion from the investigation of the second research question is that the options in multiple-choice questions can be ordered according to their place in the text in order to decrease the reading and memory load on the part of test takers so that one drawback of using multiple-choice response formats can be eliminated. Furthermore, A1 level task was found to require some modifications in terms of contextual demands of the tasks; however, the other tasks seem to meet the contextual requirements of Weir's framework and the CEFR specifications.

5.1.3 Summary of the findings for the third research question

For the investigation of the third research question, the tasks both in the first and second administrations of the test were analyzed in terms of central tendency measures, reliability and item statistics. Based on the statistical analyses of both pilot administrations, it can be seen that the reliability values decreased after the revisions made on the first version of the tasks. This situation can be attributed to the smaller number of test takers the fewer number of tasks and items in the second administration. Classical item analysis can give different results for different groups of test takers. Therefore, administering the task with a larger group of test takers once more may provide us with more reliable data. As opposed to the reliability scores, the mean scores of the items and the tasks improved and they were ordered in a more expected way after the second piloting. One serious problem was noticed about A1 level task, which had lower mean scores than A2 level task. With all the discussion up to now combined, it can be maintained that A1 level task requires serious changes in terms of contextual validity and scoring validity. Other than A1 level task, the other tasks provide satisfactory and reasonable results in terms of scoring validity.

5.2 Limitations of the study

As previously mentioned, the test was administered twice and the sample size in the second administration was quite small since we had to divide the participants into two groups due to practicality reasons. The sample size must have affected the results of the statistical analyses and therefore, led to worse reliability coefficients. Furthermore, the ability level of the test takers and its variability also affects reliability and in second administration of the test, the two groups were quite

homogeneous and had very similar language abilities. Therefore, the conclusions discussed in this study must be considered tentative. The test should be implemented again with a larger number of test takers with more heterogeneous backgrounds in order to reach more conclusive results. The number of tasks in the test can be considered as another limitation of the study. C1 level task in the first administration was discarded and a new task was not added to replace it due to time limitations. Increasing the number of tasks and items can also make positive contributions to the scoring validity of the test. Therefore, in the future versions of the test inclusion of a C1 level task can be considered. Lastly, the analyses of the tasks in terms of their linguistic features were considered to be rather incomplete since it was not possible to use various sources of text and task analysis for Turkish. The CEFR level descriptors were taken as a reference for the topics, lexical items, grammatical structures and functional languages to be included in the listening texts. However, since these descriptors were not prepared for Turkish, a comprehensive analysis of the texts was not possible. More research into the analysis of Turkish language can help test developers determine the linguistic difficulty of the listening texts in the future.

Despite these limitations and some problems that appeared as a result of the various analyses, it can still be argued that this test offers encouraging results as a newly developed listening test. The theoretical background of the study is clearly defined and explained, and a substantial amount of data is collected to support our validity claims. With the modifications mentioned above, it will develop into a more valid and reliable measure of assessing the listening skill in Turkish as a foreign language.

5.3 Suggestions for further research

In addition to the discussion above, a few more suggestions and implications for further research will be mentioned below. This study only focuses on three aspects of Weir's (2005) validation framework; theory-based, context and scoring validity. Therefore, effects of test taker characteristics, consequential validity and criterion-related validity can also be investigated in order to conduct a more comprehensive validation study. For instance, the effects of different test taker characteristics on the test scores can be examined. In addition, the positive and negative effects of the test on test takers can be explored after the administration of the test. Test takers' performances in the Turkish classes they are admitted to can be compared with their performances on the current test. These further studies can provide a better picture on the validity of the listening test under investigation.

Moreover, this research study demonstrated a very important finding: The two important facets of test development, i.e. validity and reliability, do not guarantee each other. After the first administration of the present test, although the reliability values were very good, there were concerns about the construct validity of the test and since we could not justify the operationalization of the construct, we had to make revisions on the first version of the test. Reliability, on its own, was not a decisive factor on the use of the test. After the second administration, it was observed that the construct was operationalized in a more justifiable way and therefore the construct validity claims of the test were stronger; however, the reliability measures of the test and the items were not as satisfactory as in the first analysis. This indicated that technical quality of the items also impacted on the reliability of the test despite the higher construct validity of the test. Therefore, it can be concluded that as Bachman (1990), Messick (1993) and Weir (2005) state in Chapter 2, validity and

reliability complement each other and both a priori (for validity) and a posteriori (for reliability) evidence collection is necessary for test validation. Achieving good results in only validity or reliability does not ensure the quality of a test.

A final suggestion is related with research on Turkish language. Turkish is a newly emerging language as a foreign language to be taught and assessed and therefore, research on this area is rather limited. There need to be tools or soft ware which analyze the difficulty of spoken and written texts in Turkish. Moreover, researchers need to provide more resources which demonstrate Turkish corpus data including frequent and infrequent words, idioms, concordances, collocations and so on in the Turkish language. The level of abstractness or concreteness in a text and the genre of a text should be analyzed via automatic test tools. Without these resources, it is a very challenging task to standardize the difficulty of test texts. With more advancements in this field, the quality of tests assessing Turkish as a foreign language will improve.

APPENDIX A

CEFR DESCRIPTORS FOR OVERALL LISTENING COMPREHENSION

	OVERALL LISTENING COMPREHENSION
C2	Has no difficulty in understanding any kind of spoken language, whether live or broadcast, delivered at fast native speed.
C1	Can understand enough to follow extended speech on abstract and complex topics beyond his/her own field, though he/she may need to confirm occasional details, especially if the accent is unfamiliar. Can recognise a wide range of idiomatic expressions and colloquialisms, appreciating register shifts. Can follow extended speech even when it is not clearly structured and when relationships are only implied and not signalled explicitly.
B2	Can understand standard spoken language, live or broadcast, on both familiar and unfamiliar topics normally encountered in personal, social, academic or vocational life. Only extreme background noise, inadequate discourse structure and/or idiomatic usage influences the ability to understand. Can understand the main ideas of propositionally and linguistically complex speech on both concrete and abstract topics delivered in a standard dialect, including technical discussions in his/her field of specialisation. Can follow extended speech and complex lines of argument provided the topic is reasonably familiar, and the direction of the talk is sign-posted by explicit markers.
B1	Can understand straightforward factual information about common everyday or job related topics, identifying both general messages and specific details, provided speech is clearly articulated in a generally familiar accent. Can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure etc., including short narratives.
A2	Can understand enough to be able to meet needs of a concrete type provided speech is clearly and slowly articulated. Can understand phrases and expressions related to areas of most immediate priority (e.g. very basic personal and family information, shopping, local geography, employment) provided speech is clearly and slowly articulated.
A1	Can follow speech which is very slow and carefully articulated, with long pauses for him/her to assimilate meaning.

(Source: The Council of Europe, 2001, p.66)

APPENDIX B

TASKS FOR THE FIRST PILOTING

YABANCILAR İÇİN TÜRKÇE SINAVI DİNLEDİĞİNİ ANLAMA

Bölüm 1

3 tane kısa konuşma dinleyeceksiniz. Konuşmalara göre doğru cevapları yazınız.

Konuşma 1:

1. Öğrenci kırtasiyeden neler almıştır? (İki tanesini yazınız.)

2. Öğrenci ne kadar para ödemiştir?

Konuşma 2:

3. Öğrenci belgesi başvurudan kaç gün sonra hazır oluyor?

4. Öğrenciler için kaç tane öğrenci belgesi ücretsizdir?

Konuşma 3:

5. Öğrencinin sağlık sorunu neresindedir?

6. Öğrenci merhemi günde kaç defa kullanacaktır?

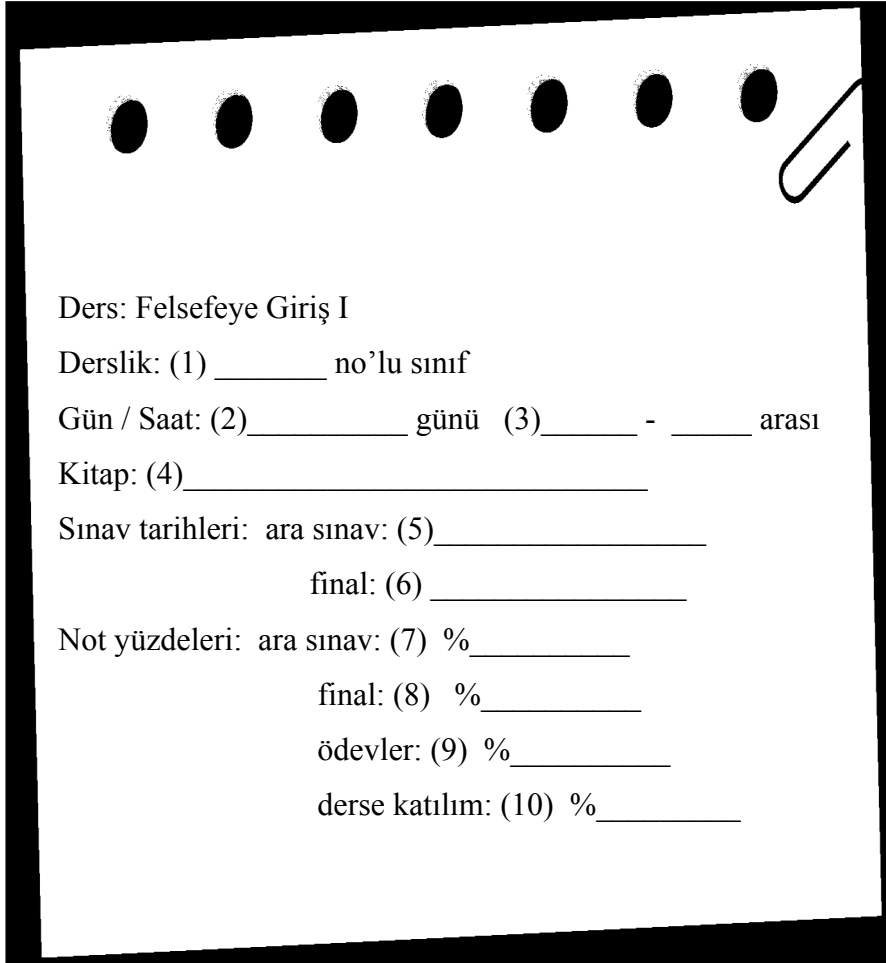
Bölüm 2

İki arkadaş arasındaki konuşmayı dinleyeceksiniz. Konuşmaya göre aşağıdaki sorulara doğru cevapları işaretleyiniz.

1. Uğur konferansa neden gitmemiştir?
 - a. İyi konuşmacılar olmadığı için
 - b. Kermese gitmek için
 - c. Sergiye gitmek için
 - d. Uyanamadığı için
2. Uğur hangi kulübün partisine gitmiştir?
 - a. Ekonomi Kulübü
 - b. Güzel Sanatlar Kulübü
 - c. Sosyal Girişimcilik Kulübü
 - d. Spor Kulübü
3. Zeynep hafta sonu kimin düğününe gitmiştir?
 - a. Ablasının
 - b. Bir akrabasının
 - c. Bir arkadaşının
 - d. Erkek kardeşinin
4. Uğur ailesini görmeye ne zaman gidecektir?
 - a. Ocak'ta
 - b. Mayıs'ta
 - c. Haziran'da
 - d. Temmuz'da
5. Uğur...
 - a. Psikoloji ödevini unutmuştur.
 - b. Sosyoloji ödevini unutmuştur.
 - c. Sosyoloji ödevini çok hızlı yapmıştır.
 - d. Sosyoloji ödevini de Psikoloji ödevini de yapmamıştır.
6. Zeynep' göre...
 - a. Hoca ek süre verecektir.
 - b. Hoca ek süre vermeyecektir.
 - c. Hocanın ek süre verip vermeyeceği kesin değildir.
 - d. Uğur kesinlikle ek süre istemelidir.

Bölüm 3

Üç arkadaş arasında geçen bir konuşmayı dinleyeceksiniz. Boşlukları konuşmaya göre doldurunuz. Boşluklara EN FAZLA İKİ kelime ve/veya sayı yazınız.



Ders: Felsefeye Giriş I

Derslik: (1) _____ no'lu sınıf

Gün / Saat: (2) _____ günü (3) _____ - _____ arası

Kitap: (4) _____

Sınav tarihleri: ara sınav: (5) _____

final: (6) _____

Not yüzdeleri: ara sınav: (7) % _____

final: (8) % _____

ödevler: (9) % _____

derse katılım: (10) % _____

Bölüm 4

Bir kurs duyurusu dinleyeceksiniz. Duyuruya göre aşağıdaki cümleleri D (Doğru), Y (Yanlış) ya da YA (Yer Almıyor) olarak işaretleyiniz.

1. Kursa okulun öğrencisi olmayan kişiler kayıt yaptırabilecektir.	D	Y	YA
2. Bilgisayar bilgisi olan kişiler başlangıç düzeyindeki kursa kayıt yaptıramayacaklardır.	D	Y	YA
3. Başlangıç düzey kursun katılımcıları iki tane sınava gireceklerdir.	D	Y	YA
4. Öğrenciler kursa kişisel bilgisayarlarıyla gelmek zorundadır.	D	Y	YA
5. Orta düzey kursun katılımcıları kurs sonunda bir proje yarışmasına katılacaklardır.	D	Y	YA
6. Kurs Ağustos ayında bitecektir.	D	Y	YA
7. Dersler üniversitedeki akademisyenler tarafından verilecektir.	D	Y	YA
8. Öğrenciler kurs için 400 TL ücret ödeyeceklerdir.	D	Y	YA
9. Öğrenciler kayıt ücretini nakit ya da taksitle ödeyebileceklerdir.	D	Y	YA
10. Kurs duyurusu Teknoloji Birimi'nden gelen bir görevli tarafından yapılmıştır.	D	Y	YA

Bölüm 5

Temel Bilgi Teknolojileri dersinden bir kesit dinleyeceksiniz. Dinlediğiniz derse göre aşağıdaki boşlukları EN FAZLA İKİ kelime ile doldurunuz.

Bilgi Teknolojileri I

21/10/2014

- Ağ: Farklı bilgisayarların birbirleriyle fiziksel olarak iletişim kurabildiği ortamlar

- Ağ kurabilmek için gerekenler:

1. İki farklı (1)_____

2. İnternete girebilmeleri için onlara ait ağ (2) _____ ve ağ kartları

- Coğrafi alan kriterlerine göre ağlar:

1. Yerel alan ağları

Bir (3)_____ içerisinde kullanılır. Kablo uzunluğu en fazla 100 metredir. Diğer ağlara göre daha (4)_____.

2. Metropolitan alan ağları

(5)_____ ağlarından oluşur.

3. Geniş alan ağları

(6)_____ olarak farklı yerlerdeki bilgisayarların birbirine bağlanmasıyla oluşur. (Örnek: (7)_____)

- Ağ topolojisi: Kablolu ağlarda bilgisayarların birbirlerine bağlanma şeklidir.

1. Veri Yolu topolojisi

İlk topolojidir. Tüm bilgisayarlar tek bir kabloyla birbirlerine bağlanır.

Kullanımı daha kolay ve ucuzdur, ama performansı bazen (8)_____.

2. Halka topolojisi

Bilgisayarlar halka şeklinde yerleştirilir. Her bilgisayarın iki

9)_____ olur.

3. Yıldız topolojisi

(10)_____ kullanılan topolojidir. Her bilgisayar tek bir kabloyla göbek adı verilen bir cihaza bağlanır.

A1 LEVEL LISTENING TEXT FOR THE FIRST PILOTING

Konuřma 1:

MÜřTERİ: Merhaba.

KIRTASIYECİ: Merhaba, buyurun.

MÜřTERİ: Ben dosya almak istiyorum da, ne çeřitleriniz var?

KIRTASIYECİ: Elimizde řu anda sunum dosyaları var. Sert kapaklı, içinde 20 tane bölmesi var. Bir de tekli řeffaf dosyalarımız var. Hangisini vereyim?

MÜřTERİ: Sunum dosyalarının tanesi ne kadar?

KIRTASIYECİ: 3.5 lira.

MÜřTERİ: Peki, řeffaf dosyalar ne kadar peki?

KIRTASIYECİ: Tanesi 15 kuruř.

MÜřTERİ: Beř tane řeffaf dosya alayım o zaman.

KIRTASIYECİ: Buyurun.

MÜřTERİ: řu tükenmez kalemle not defterinin fiyatı nedir peki?

KIRTASIYECİ: Kalem 2.25. Not defteri de 4.75.

MÜřTERİ: Tamam. Onları da ekleyebilir misiniz?

KIRTASIYECİ: Tabi ki.

MÜřTERİ: Borcum ne kadar?

KIRTASIYECİ: Hepsi yedi lira 75 kuruř.

MÜřTERİ: Buyurun.

KIRTASIYECİ: Teřekkürler. İyi günler.

MÜřTERİ: İyi günler, kolay gelsin.

Konuřma 2:

ÖĞRENCİ: Merhaba, öğrenci işleri burası mı?

ÖĞRENCİ İřLERİ MEMURU: Evet burası. Nasıl yardımcı olabilirim?

ÖĞRENCİ: Ben öğrenci belgesi almak istiyorum. Ne yapmalıyım?

ÖĞRENCİ İŞLERİ MEMURU: Öğrenci belgesini hemen veremiyoruz. Önce internetten başvuru yapmalısınız. Başvurunun ertesi günü belgeniz hazır olur. Gelip buradan alabilirsiniz.

ÖĞRENCİ: Peki, belge ücreti nedir?

ÖĞRENCİ İŞLERİ MEMURU: Ücret ödemenize gerek yok. Beş adet belge ücretsiz.

Beşten sonrası için bir lira 50 kuruş ödemeniz gerekiyor.

ÖĞRENCİ: Harika, teşekkürler bilgi için. Kolay gelsin.

ÖĞRENCİ İŞLERİ MEMURU: Sağ olun, iyi günler.

Konuşma 3:

DOKTOR: Merhaba, hoş geldiniz.

HASTA: Merhaba, benim bir problemim vardı da...

DOKTOR: Buyurun oturun....Sorun nedir?

HASTA: Kolumla ilgili ufak bir kaza geçirdim. Yurtta oda kapısının arkasında duruyordum. Arkadaşım birden açtı kapıyı, ben de görmedim. Kapı hızlıca dirseğime çarptı. Kırık olabilir mi?

DOKTOR: Bir bakalım...Kolunuzu hareket ettirebiliyor musunuz?

HASTA: Evet.

DOKTOR: Dirseğinizi oynatabiliyor musunuz?

HASTA: Evet, ama çok az.

DOKTOR: Anladım. Kolunuz kırık değil. Dirseğinizi oynatıyorsunuz, ama ezilmiş. Şu sargı bandını saralım kolunuza. Size bir de merhem yazıyorum. Sabah akşam sürün lütfen. Ağrınız olur diye de bir ağrı kesici ilaç yazıyorum. Eğer kendinizi kötü hissederseniz, kolunuzu hareket ettiremezseniz, tekrar gelin lütfen. Film çektirmemiz gerekebilir.

HASTA: Çok teşekkürler. Kolay gelsin, iyi günler.

A2 LEVEL LISTENING TEXT FOR THE FIRST PILOTING

ZEYNEP: Merhaba Uğur, günaydın, n'aber?

UĞUR: İyilik Zeynep, senden n'aber?

ZEYNEP: Ben de iyiyim. Hafta sonun nasıldı?

UĞUR: Baya yoğundu ama eğlenceliydi. Bir sürü yere gittim. Cumartesi sabah Ekonomi Kulübü'nün konferansına gidecektim, çok önemli konuşmacılar vardı ama sabah alarmı duymamışım, kaçırdım konferansı. Sonra biz de arkadaşlarla Sosyal Girişimcilik Kulübü'nün kermesine gittik. Orada bir şeyler yiyip Güzel Sanatlar Kulübü'nün sergisine geçtik. Sergide baya ilginç resimler vardı, vaktin olursa kesin git bak... Neyse, akşam da Spor Kulübü'nden arkadaşlar aradı. Yeni üyelere hoş geldin partisi veriyorlarmış, biz de oraya gittik. Çok eğlendik. Bir sürü yeni insanla tanıştık. Pazar günü de bütün gün dinlendim, Cumartesi çok yorulmuşum. Senin hafta sonun nasıldı, sen n'aptın?

ZEYNEP: Benim hafta sonum da çok yorucuydu ama çok güzel geçti. Ben de hafta sonu İzmir'deydim. Hatta yurttan bir arkadaşımla gittik. Amcamın kızı evlendi bu hafta sonu, o yüzden gittim zaten. Erkek kardeşim hariç aileden herkesi gördüm. Kardeşimin sınavları vardı, okuldaydı o, ama ablamla baya vakit geçirdik. Çoğu akrabamı da gördüm. Düğün çok güzel geçti, baya stres atmış oldum.

UĞUR: Ne güzel. Ben de ailemi çok özledim. Ocak ayında evde fazla kalamadım, arkadaşlarla tatil planı yapmıştık. Mayıs ayında gitmek istiyordum ama o zaman da sınavlar var. Durum böyle olunca ben de Haziran'a bilet aldım, finallerden sonra giderim. Zaten Temmuz'da stajım var, yine bütün yaz buradayım. Neyse, hafta sonu çok gezdik, ödevleri aksattık sanırım biraz. Sen Sosyoloji ödevini yaptın mı?

ZEYNEP: Evet, düğüne gitmeden önce yetiştirdim ama çok hızlı yazdım, biraz aceleye geldi sanırım. Umarım hoca beğenir. Sen yazdın mı peki?

UĞUR: Sen en azından hızlı olsa da yapmışsın. Ben hep unutmuşum. Psikoloji ödevini ancak yapabildim zaten. Sosyoloji tamamen aklımdan çıkmış. Bu sabah aklıma geldi. Aslında hiç yapasım yok ama yapmam lazım yoksa dersten kalacağım. Bugün ödev için hocadan ek süre isteyeceğim. Umarım verir. Geçen dönem vermişti, buna da verir herhalde değil mi?

ZEYNEP: Ya evet geçen dönem bir kere ek süre vermişti ama genelde vermiyor diyorlar. Bilmiyorum ki... Ruh haline bağlı biliyorsun. Eğer kızgın görünüyorsa bir şey söyleme sakın. Ama neşesi yerindeyse konuşabilirsin, yani. Verirse iyi olur aslında ben de ödevimin üstünden geçerim.

UĞUR: Neyse çıkışta konuşuruz, hadi hoşça kal.

ZEYNEP: Hadi görüşürüz, bye bye!

B1 LEVEL LISTENING TEXT FOR THE FIRST PILOTING

ÖĞRENCİ 1: Selam, n'aber?

ÖĞRENCİ 2: İyilik, sizden n'aber?

ÖĞRENCİ 1: İyiyiz ya merak ettik seni, n'oldu? Derse niye gelmedin? Sen dersleri hiç kaçırmazsın. Hele ilk dersleri hiç kaçırmazsın. Hocalar bir sürü şey anlatıyor sonuçta, sınav, ödev, tarih falan.

ÖĞRENCİ 2: Evet haklısın, ama sabah midem ağrıyordu, doktora gitmek zorunda kaldım. Gelemedim. Hoca neler anlattı söylesenize? Not alayım ben şuraya.

ÖĞRENCİ 3: Hoca genel bilgiler verdi işte ders yeri, saatleri, sınavlar, sınav tarihleri, ödevler falan.

ÖĞRENCİ 2: Bu dersin adı Felsefeye Giriş 1'di değil mi?

ÖĞRENCİ 3: Evet.

ÖĞRENCİ 2: Tamam. Peki, dersler ne zaman nerede olacakmış? Kayıt sayfasına baktım ama hala belli değildi.

ÖĞRENCİ 1: Evet orada görünmüyormuş ama hoca bilgi verdi. Ders haftada üç saatmiş, hepsi peş peşe zaten. Çarşamba günleri 1'den 4'e kadar olacakmış.

ÖĞRENCİ 2: Peki dersler nerede yapılacaktı?

ÖĞRENCİ 3: Bu haftaki e-postasında 116 numaralı sınıfa gelin demişti ama normalde 106 numaralı sınıfta olacakmış.

ÖĞRENCİ 2: Tamam, 106 yazıyorum o zaman. Kitap hakkında bir şey dedi mi hoca?

ÖĞRENCİ 3: Evet. Felsefenin İlkeleri diye bir kitap. Kuzey Kampüs'teki kitapevinde varmış. Oradan alabilirsiniz dedi.

ÖĞRENCİ 2: Eveet, anladım, Felsefenin İlkeleri, kitapevinde. Peki sınavlar nasıl olacakmış? O konuda bir şey söyledi mi?

ÖĞRENCİ 1: Evet evet, sınav tarihlerini verdi. Toplamda bi' tane ara sınav bi' tane de final olacakmış. Normalde ara sınavın tarihi 20 Nisan'dı ama tam bahar tatilinden sonra olduğu için hocayla konuştuk, öne çektik sınav tarihini.

ÖĞRENCİ 2: Aa, çok iyi olmuş.

ÖĞRENCİ 1: Evet, yani, ara sınavı 8 Nisan'da yapacak. Final tarihini de söyledi ama 6 Haziran mı dedi 9 Haziran mı dedi, tam hatırlayamıyorum. Ya şuraya not almıştım, bi bakayım. Hah, tamam 9 Haziran'daymış. Finaller 10'unda bitiyor zaten. Bu sene erken bitecek gibi.

ÖĞRENCİ 2: Evet, ama daha şimdiden stres başladı, bakalım nasıl geçecek bu dönem... Ödevler nasılmış peki, onları da not alayım...

ÖĞRENCİ 3: Bazen kısa ödevler bazen de makale yazacakmışız. Onları zamanı geldiğinde söyleyeceğim dedi. Ama toplamda 1 tane makale var sanırım. Ödevler notumuzun %20'sini etkileyecekmiş.

ÖĞRENCİ 2: Öyle mi? Diğer şeylerin yüzdesi neymiş, sınavların falan?

ÖĞRENCİ 3: Ara sınav %30 dedi. Baya bir yüksek yüzdesi, neredeyse final kadar. Finalin de %40'mış.

ÖĞRENCİ 2: Kalan %10'luk kısım ne peki?

ÖĞRENCİ 1: Derse katılım, yani. Okumalarınızı yapmadan gelmeyin, ders devamlı tartışma halinde geçecek dedi hoca. Yani toplamda iki sınav var ama ödevler, okumalar falan, baya yoğun geçecek bi' ders.

ÖĞRENCİ 2: Evet öyle görünüyor. Bunları öğrendiğim iyi oldu kızlar, çok sağ olun. Neyse ben kaçayım, daha Edebiyat dersinin notlarını alacağım. Haftaya derste görüşürüz o zaman.

ÖĞRENCİ 3: Tamam görüşürüz, hadi kolay gelsin sana. Güle güle.

ÖĞRENCİ 1: Hadi görüşürüz.

B2 LEVEL LISTENING TEXT FOR THE FIRST PILOTING

Merhaba arkadaşlar, günaydın! Bugün dersimize başlamadan önce size bir duyuru yapmak istiyorum. Üniversitemizin Teknoloji Birimi hem öğrencilerimize hem de dışarıdan gelmek isteyenlere açık olan sertifikalı bir bilgisayar kursu başlatacak. Kursta hem başlangıç düzeyinde hem de orta düzeyde dersler verilecek. Eğer hiç bilgisayar öğrenme ve kullanma deneyiminiz olmadıysa ya da basit işlemleri gerçekleştirebilecek kadar bir deneyiminiz var ama yine de programları çok etkili kullanamıyorsanız başlangıç düzeyindeki dersler sizin için çok uygun olacaktır. Bu düzeyde katılımcılar bilgisayar kullanımı ile ilgili temel beceriler edinecekler ve basit ama kullanımı yaygın programları etkin bir şekilde kullanmayı öğrenecekler. Kurs sonunda katılımcılar bir final sınavına girecekler. Orta düzey dersler ise evde ya da işte bir şekilde bilgisayar kullanan ancak kendini bu konuda daha da geliştirmek isteyen, daha detaylı bilgi almak isteyen kişiler için gayet uygun. Bu seviyede daha üst düzey programların kullanımına odaklanılacak ve kurs sonunda katılımcılardan özgün bir proje teslim etmeleri beklenecektir. Kurs 15 Mayıs'ta başlayacak ve toplamda 8 hafta sürecek. Dersler Pazartesi günleri 18:00-20:00 saatleri arasında olacak. Derslerin nerede yapılacağı ise henüz belli değil, o yüzden size daha sonra bildirilecek. Dersler okulumuz akademik kadrosu tarafından verilecek. Kayıtlar 3 Mayıs'ta başlayıp 5 Mayıs'ta son bulacak. Sınıf mevcutları 25 kişiyi geçmeyeceğinden katılmak isteyenlerin bir an önce kayıt yaptırmaları gerekmekte. Kayıt ücreti dört taksitte alınacak olup her taksit 50 liradır. Kayıt ücretlerini teknoloji birimindeki yetkili kişilere vermeniz gerekmektedir. Yani, şimdilik size söyleyebileceklerim bu kadar. Eğer daha fazla sorunuz olursa Teknoloji Birimi'ndeki yetkili kişilerle görüşebilirsiniz. Evet, şimdi dersimize dönecek olursak, geçen hafta...

C1 LEVEL LISTENING TEXT FOR THE FIRST PILOTING

ÖĞRETMEN: Merhaba arkadaşlar! Temel Bilgi Teknolojileri 1 programının bu son bölümünde sizlerle tekrar birlikteyiz. Bugüne kadarki programlarda konunun uzmanlarıyla birlikte uygulamalı bir şekilde konuları işlemeye çalıştık. Bugünkü konumuz ağ teknolojileri ve kablosuz ağlar. Bugünkü konuğumuz bir uzman değil, sizlerden birisi. Anadolu Üniversitesi Bilgisayar Bölümü 4. sınıf öğrencisi Ceyda Özsoy. Hoş geldin Ceyda.

ÖĞRENCİ: Hoş bulduk hocam.

ÖĞRETMEN: Ceyda istersen ağ kavramından başlayalım önce. Ağ nedir?

ÖĞRENCİ: Ağ aslında bugüne kadar hiç yabancı olmadığımız bir kavram. Eskiden, eski teknolojiye bakarsak bilgi paylaşımı, veri aktarımı için disketleri kullanıyorduk. Fakat bugün bu işi, aynı işlemleri internet aracılığıyla yapabiliyoruz. Yani ağ kavramı farklı bilgisayarların fiziksel olarak birbirlerine bağlanmasıyla oluşan bir ortamdır.

ÖĞRETMEN: Bu ortamda, bu ortamın, ağ teknolojileri dediğimiz ortamın temelinde neler var?

ÖĞRENCİ: Ağ teknolojileri dersek aklımıza iki farklı bilgisayar ve internete bağlanabilmeleri için her bilgisayara ait ağ yazılımı ve ağ kartları gerekir. Bu sayede bilgisayarlar birbirleriyle iletişim kurar.

ÖĞRETMEN: Ceyda ağın çeşitleri var mı?

ÖĞRENCİ: Evet, ağlar coğrafi alan kriterlerine göre üçe ayrılırlar. Bunlardan ilki, yerel alan ağları, ikincisi metropolitan alan ağları ve üçüncüsü ise geniş alan ağlarıdır.

ÖĞRETMEN: Ceyda bu kavramları biraz açabilir misin? Nedir bunlar?

ÖĞRENCİ: Az önce söylediğimiz gibi üç tip ağ çeşidi bulunmaktadır. Bunlardan ilki yerel alan ağları olup bir yerleşke içerisinde kullanılır. Genelde kabloların uzunluğu 100 metreyi geçmez ve diğerlerine göre daha hızlı çalışırlar. İkincisi metropolitan alan ağları olmakla birlikte yerel alan ağlarının bir araya gelmesiyle oluşan ağ çeşididir. Son olarak geniş alan ağlarından bahsetmek gerekirse, coğrafi olarak çok farklı konumlardaki bilgisayarların birbirine bağlanmasıyla oluşan ağ çeşididir. Buna verebileceğimiz en güzel örneğe dünyayı kapsayan www, yani internettir.

ÖĞRETMEN: Ceyda, kablolu ağlardan bahsederken ağdaki bilgisayarların birbirlerine nasıl bağlandıklarının da bir önemi olsa gerek. Bildiğim kadarıyla buna biz ağ topolojileri diyoruz. Ne çeşit ağ topolojileri var?

ÖĞRENCİ: Kablolu ağlarda üç çeşit topolojiden bahsetmek mümkündür. Bunları sırasıyla sayarsak, veri yolu, halka ve yıldız topolojileri. Veri yolu topolojisinden bahsetmek gerekirse, ilk kullanılan topolojidir ve tüm bilgisayarlar tek bir kablo yardımıyla birbirine bağlanır. Kablonun uzunluğu 100 metreden uzun olmamalıdır. Diğer topolojilere göre kullanımı kolay ve ucuzdur; fakat yoğun trafikte performansı daha kısıtlıdır. Halka topolojisinde bilgisayarlar halka şeklinde konumlandırılır. Yani her bilgisayarın iki komşusu olur. Çalışma şeklinden bahsedersek, gelen mesaj bilgisayarlar arasında andaç yardımıyla iletilir. Eğer bilgisayar mesajı kendisiyle ilgiliyse alır, değilse komşu bilgisayara gönderir. Bu şekilde mesaj ilgili bilgisayara gidene dek çalışma devam eder.

ÖĞRETMEN: Peki yıldız topolojisinin diğerlerinden farkı nelerdir?

ÖĞRENCİ: Yıldız topolojisi günümüzde en sık kullanılan topoloji çeşididir. Çalışma şekli her bilgisayar tek kablo ile switch yani anahtar ya da hap yani göbek adı verilen cihaza tek tek bağlanır. Mesaj iletimi ise bu şekilde switch ya da hap sayesinde sağlanır.

ÖĞRETMEN: Örnekleme gerekirse diyelim benim 20 bilgisayarlı bir şirketim var,
burada bir ağ kurmalı mıyım, kurmam gerekir mi?

ÖĞRENCİ: Örneğinizin üzerinden anlatmaya çalışırsak...

APPENDIX C

TEST TASKS FOR THE SECOND PILOTING

A1 SEVİYESİ DİNLEME SINAVI

Şimdi 3 tane kısa konuşma dinleyeceksiniz. Konuşmaları bir kere dinleyeceksiniz. Önce soruları 1 dakika içinde okuyunuz. Doğru cevapları boşluklara yazınız.

Konuşma 1: Kırtasiyede

1. Öğrenci neler satın aldı? Listeyi tamamlayınız.

- Dosya
- _____
- Not defteri

2. Öğrenci kaç lira ödedi?

Konuşma 2: Öğrenci İşleri Ofisi'nde

3. Öğrenci belgesi kaç günde hazır oluyor?

4. Kaç tane öğrenci belgesi ücretsiz?

Konuşma 3: Doktorda

5. Öğrencinin neresinde sorun var?

6. Doktor öğrenciye ne verdi?

A2 SEVİYESİ DİNLEME SINAVI
FELSEFEYE GİRİŞ DERSİ NOTLARI

Şimdi iki arkadaş arasında yurttan geçen bir konuşmayı dinleyeceksiniz. Konuşmayı bir kere dinleyeceksiniz. Önce soruları 1 dakika içinde okuyunuz. Doğru cevapları işaretleyiniz ya da boşluklara yazınız.

1. Ayşe neden derse gelemedi? Doğru cevabı seçiniz.
 - a. Başka dersin ödevi vardı.
 - b. Sağlık sorunu vardı.
 - c. Dersin saatini şaşırdı.
2. Ders haftanın hangi günü olacak? _____
3. Ders _____ - _____ saatleri arasında olacak.
4. Ders kitabının yazarının soyadı ne? _____
5. 8 Nisan'daki ara sınav hangi tatilden önce?

6. Final sınavı hangi dersin finalinden sonra?

Dersin ödevleri neler? Aşağıya yazınız.

7. _____
8. _____

B1 SEVİYESİ DİNLEME SINAVI

HAFTA SONU NE YAPTIN?

Şimdi Zeynep ve Uğur adlı iki arkadaş arasında geçen hafta sonu etkinlikleriyle ilgili bir konuşmayı dinleyeceksiniz. Konuşmayı bir kere dinleyeceksiniz. Önce soruları 1 dakika içinde okuyunuz. Doğru cevapları işaretleyiniz.

1. Uğur konferansa neden gitmedi?
 - a. Ali'yle sinemaya gittiği için
 - b. Eğlenmeye gittiği için
 - c. Uyanamadığı için
 - d. Yoğun olduğu için
2. Uğur cumartesi akşamı hangi kulübün etkinliğine gitti?
 - a. Ekonomi Kulübü
 - b. Güzel Sanatlar Kulübü
 - c. Sosyal Girişimcilik Kulübü
 - d. Spor Kulübü
3. Zeynep hafta sonu kimin düğününe gitti?
 - a. Ablasının
 - b. Bir akrabasının
 - c. Bir arkadaşının
 - d. Erkek kardeşinin
4. Uğur ailesini görmeye ne zaman gidecek?
 - a. Ocak'ta
 - b. Mayıs'ta
 - c. Haziran'da
 - d. Temmuz'da
5. Uğur Sosyoloji ödevini...
 - a. çok hızlı yaptı.
 - b. hiç yapmayacak.
 - c. yapmayı unuttu.
 - d. zamanında yetiştirdi.
6. Zeynep hocanın ek süre...
 - a. vereceğinden emindir.
 - b. vermeyeceğinden emindir.
 - c. verip vermeyeceği konusunda kararsızdır.
 - d. vermesine karşıdır.

B2 SEVİYESİ DİNLEME SINAVI

RADYO PROGRAMI

Şimdi “Bir Varmış Bir Yokmuş” Masal Şenliği hakkında bir radyo programı dinleyeceksiniz. Programı bir kere dinleyeceksiniz. Soruları 3 dakika içinde okuyunuz. Doğru cevapları işaretleyiniz.

1. Hangisi masal şenliğinin bir özelliğidir?
 - a. Belediye tek başına organize etmiştir.
 - b. Bu yıl üçüncüsü yapılacaktır.
 - c. Türkiye’de başka benzeri yoktur.
 - d. Yeni popüler olmaya başlamıştır.
2. Etkinlik koordinatörü şenlik hakkında ne düşünüyor?
 - a. Şenliğin uluslararası olması özellikle önemlidir.
 - b. Belli yaştaki insanlara hitap etmektedir.
 - c. Park şenlik için güvenli bir yerdir.
 - d. Büyük etki uyandıracaktır.
3. Yetişkinler hangi masal anlatım etkinliğinde rol alabilecek?
 - a. Dansla masal anlatımı
 - b. Müzikle masal anlatımı
 - c. Doğaçlama masal anlatımı
 - d. Pantomimle masal anlatımı
4. Müzeler şenliğe neden katılacak?
 - a. Çocuklara tarihi eserleri göstermek için
 - b. Çocukları sanatla tanıştırma yolu olduğu için
 - c. Şenlikte atölye çalışmaları gerektiği için
 - d. Yurtdışında çok yaygın bir uygulama olduğu için
5. Şenlik programıyla ilgili hangisi doğrudur?
 - a. Amatör sanatçılar da şenliğe katılacaktır.
 - b. Çocuklar için pek çok masal çeşidi olacaktır.
 - c. Çocuklara yönelik bütün gün etkinlikler vardır.
 - d. Şenlik akşam 6’dan sonra bitmektedir.
6. Şenlik geçidiyle ilgili hangisi yanlıştır?
 - a. Katılımcıların tamamı İstanbul’dan olacak.
 - b. Bir devlet adamı bir konuşma yapacak.
 - c. Şenlik geçidi öğleden önce olacak.
 - d. Şenlik geçitten sonra başlayacak.

7. Konuşmacıya göre şenliğin toplumsal amacı nedir?

- a. Toplumun yaratıcılığını geliştirmek
- b. Toplumun kaygısızca zaman geçirmesini sağlamak
- c. Toplumun çağdaşlaşmasına yardımcı olmak
- d. Toplumsal bağlarımızın güçlenmesini sağlamak

8. soruyu dinleme bittikten sonra cevaplayınız.

8. Konuşmadan Masal Şenliği ile ilgili hangisi çıkarılabilir?

- a. Masalların çağdaşlaşmasına katkıda bulunacak.
- b. Toplumun eğitimine katkıda bulunacak.
- c. Türk kültüründe önemli bir öğeyi destekleyecek.
- d. Türk masallarını diğer uluslara tanıttak.

A1 LEVEL LISTENING TEXT FOR THE SECOND PILOTING

Konuřma 1:

MÜřTERİ: Merhaba.

KIRTASIYECİ: Merhaba, buyurun.

MÜřTERİ: Ben dosya almak istiyorum. Ne çeřitleriniz var?

KIRTASIYECİ: Çok güzel sunum dosyalarımız var. Plastik kapaklı... Renk çeřidi çok. İçlerinde 20 tane bölme var. Bir de ince, tekli dosyalarımız var. Hangisini vereyim?

MÜřTERİ: Sunum dosyalarının tanesi ne kadar?

KIRTASIYECİ: 3,5 lira.

MÜřTERİ: Peki, tekli dosyalar ne kadar?

KIRTASIYECİ: Tanesi 15 kuruř.

MÜřTERİ: Beř tane tekli dosya alayım o zaman.

KIRTASIYECİ: Buyurun.

MÜřTERİ: řu tükenmez kalemle not defterinin fiyatı nedir peki?

KIRTASIYECİ: Kalem iki lira 25 kuruř, not defteri de dört lira 75 kuruř.

MÜřTERİ: Tamam. Onları da ekleyebilir misiniz?

KIRTASIYECİ: Tabi ki.

MÜřTERİ: Borcum ne kadar?

KIRTASIYECİ: Hepsi yedi lira 75 kuruř.

MÜřTERİ: Buyurun.

KIRTASIYECİ: Teřekkürler. İyi günler.

MÜřTERİ: İyi günler, kolay gelsin.

Konuřma 2:

ÖĐRENCİ: Merhaba, öĐrenci iřleri ofisi burası mı?

ÖĐRENCİ İřLERİ MEMURU: Evet burası. Nasıl yardımcı olabilirim?

ÖĐRENCİ: Ben öĐrenci belgesi almak istiyorum. Ne yapmalıyım?

ÖĐRENCİ İřLERİ MEMURU: Maalesef öĐrenci belgesini hemen veremiyoruz.

Önce internetten başvuru yapmalısınız. Başvurudan iki gün sonra belgenizi buradan alabilirsiniz.

ÖĐRENCİ: Peki, belge ücreti nedir?

ÖĐRENCİ İřLERİ MEMURU: Ücret ödemenize gerek yok. Beř adet belge ücretsiz.

Beř taneden sonra iki lira ödemeniz gerekiyor.

ÖĐRENCİ: Harika, teřekkürler bilgi için. Kolay gelsin.

ÖĐRENCİ İřLERİ MEMURU: İyi günler.

Konuřma 3:

DOKTOR: Merhaba, hoř geldiniz.

HASTA: Merhaba, benim bir problemim vardı da...

DOKTOR: Buyurun oturun....řikayetiniz nedir?

HASTA: Yurtta bir kaza geçirdim. Oda kapısının arkasında duruyordum, arkadaşım birden kapıyı açtı, ben de görmedim, kapı hızlı bir řekilde koluma ve başıma çarptı.

DOKTOR: Bir bakalım...Kolunuzu hareket ettirebiliyor musunuz?

HASTA: Evet.

DOKTOR: DirseĐinizi oynatabiliyor musunuz?

HASTA: Evet.

DOKTOR: Kolunuzda bir sorun yok. Bir de başınıza bakalım.

HASTA: Aaahhhhh!

DOKTOR: Anladım. Evet, başınızın arkası biraz şişmiş. Önemli değil. Üzerine biraz buz koyalım. Daha sonra başınız ağrırsa diye size ağrı kesici bir ilaç veriyorum. Bu ilaçtan günde üç tane içebilirsiniz. Eğer kendinizi kötü hissederseniz, tekrar gelin lütfen. Gerekirse röntgen çektirirsiniz.

HASTA: Çok teşekkürler. Kolay gelsin, iyi günler.

DOKTOR: Geçmiş olsun.

A2 LEVEL LISTENING TEXT FOR THE SECOND PILOTING

ÖĞRENCİ 1: Selam Ayşe, n'aber?

ÖĞRENCİ 2: İyilik senden n'aber?

ÖĞRENCİ 1: İyiyim. Merak ettim seni, n'oldu? Felsefeye Giriş dersine niye gelmedin? Sen dersleri hiç kaçırmazsın. Hele ilk dersleri hiç kaçırmazsın. Hocalar bir sürü şey anlatıyor sonuçta, sınavlar, ödevler, tarihler falan.

ÖĞRENCİ 2: Evet haklısın ama sabah midem ağrıdı, doktora gittim. Gelemedim. Hocalar neler anlattı söylesene? Not alayım ben de şuraya.

ÖĞRENCİ 1: Hoca genel bilgiler verdi işte ders yeri, ders saati, sınavlar, ödevler falan.

ÖĞRENCİ 2: Tamam. Peki, ders ne zaman yapılıyor? İnternetten baktım ama yoktu.

ÖĞRENCİ 1: Evet orada yokmuş, hoca bilgi verdi. Çarşamba günleri Güney Kampüs'te olacakmış. Haftada üç saatmiş, hepsi peş peşe yapılacaktı ama yirmi dakikalık bir ara olacakmış. Yani dersler 1'de başlayıp 4'te bitecek.

ÖĞRENCİ 2: Anladım, peki dersler nerede yapılacaktı?

ÖĞRENCİ 1: Önce 206 no'lu sınıf dedi ama sonra değiştirdi. Dersler 106 numaralı sınıfta yapılacak.

ÖĞRENCİ 2: Tamam, 106 yazıyorum o zaman. Kitapla ilgili bir şey dedi mi hoca?

ÖĞRENCİ 1: Evet. Felsefeye Giriş diye bir kitap. Kuzey Kampüs'teki kitapevinde varmış. Kitabın yazarı Ahmet Aslan.

ÖĞRENCİ 2: Ahmet ne?

ÖĞRENCİ 1: Ahmet Aslan.

ÖĞRENCİ 2: Anladım, tamam, tamam. Peki ya sınavlar? Sınavlar konusunda bir şey söyledi mi?

ÖĞRENCİ 1: Evet, sınav tarihlerini verdi. Bir tane ara sınav bir tane de final sınavı olacak. Ara sınavı 8 Nisan'da yapacaktı. Yani sınav bahar tatilinden önce olacak.

ÖĞRENCİ 2: Aa, ne güzel. Bahar tatilinde dinleniriz o zaman.

ÖĞRENCİ 1: Evet bence de güzel olacak. Aslında final tarihini de söyledi ama 6 Haziran mı 9 Haziran mı dedi, tam hatırlamıyorum, bi' dakika, şuraya not almıştım, bi' bakayım. (...) Hah, tamam, 9 Haziran'mış. 6 Haziran'da Tarih finali var. Felsefe finali Tarih'ten sonra olacaktı.

ÖĞRENCİ 2: Tamam 9 Haziran yazıyorum o zaman. Ödevler nasılmış peki, onları da not alayım...

ÖĞRENCİ 1: Ödevimiz çok yok sanırım. Bir tane makale yazacağız, bir tane de sunum yapacağız.

ÖĞRENCİ 2: Peki, makale uzun mu olacak kısa mı?

ÖĞRENCİ 1: Çok uzun değil, sanırım üç-beş sayfa arası.

ÖĞRENCİ 2: Araştırma projesi de yapacak mıyız?

ÖĞRENCİ 1: Öyle bir şeyden bahsetmedi hoca.

ÖĞRENCİ 2: Peki sunum kaç dakika olacak?

ÖĞRENCİ 1: 15-20 dakika sürecektir.

ÖĞRENCİ 2: Tamam, ya çok teşekkür ederim gerçekten çok sağ ol.

ÖĞRENCİ 1: Haftaya görüşürüz o zaman.

ÖĞRENCİ 2: Görüşürüz!

B1 LEVEL LISTENING TEXT FOR THE SECOND PILOTING

ZEYNEP: Merhaba Uğur, günaydın, n'aber?

UĞUR: İyi Zeynep, sen nasılsın?

ZEYNEP: İyilik, n'olsun... Hafta sonun nasıldı?

UĞUR: Baya yoğundu ama eğlenceliydi. Aslında Ali'yle cumartesi günü sinemaya gidelim demiştik. Sonra ben fikrimi değiştirdim, Ekonomi Kulübü'nün konferansına gitmeye karar verdim, ama sabah telefonun alarmını duymamışım, kaçırdım konferansı.

ZEYNEP: Aaa, hadi ya. Kötü olmuş.

UĞUR: Evet ya, ben de çok üzüldüm ama n'apalım. Öyle olunca biz de arkadaşlarla Sosyal Girişimcilik Kulübü'nün kermesine gittik. Kermeste çok güzel yiyecekler vardı. Orada hep birlikte kahvaltı ettik. Sonra da saat 12 gibi Güzel Sanatlar Kulübü'nün sergisine geçtik.

ZEYNEP: Öyle mi? Sergi nasıldı peki?

UĞUR: Ben çok beğendim sergiyi. Baya ilginç resimler vardı, vaktin olursa kesinlikle gitmelisin... Biz öğlen gittik ama akşama kadar açık sanırım.

ZEYNEP: Hımm, ben de gideyim bakalım. Eee, sergiden sonra ne yaptınız?

UĞUR: Akşam da Spor Kulübü'nden arkadaşlar aradı. Yeni üyelere hoş geldin partisi yapacaklarını söylediler. Ben de oraya gittim. Çok eğlendik. Bir sürü yeni insanla tanıştık. Öyle işte, cumartesi günü yorulduğum için pazar günü bütün gün dinlendim. Senin hafta sonun nasıldı, sen n'aptın?

ZEYNEP: Benim hafta sonum da çok yorucu olmasına rağmen çok güzel geçti.

Cuma sabahtan İzmir'e gittim.

UĞUR: Hadi ya, neden?

ZEYNEP: Kuzenimin düğünü vardı bu hafta sonu, o yüzden gittim. Herkes gelmişti düğüne. Yaa anneleri falan acayip özlemişim. Ablamla falan da baya vakit geçirdik. Çoğu akrabamı da gördüm düğünde. Yani her şey çok güzel geçti, baya stres attım.

UĞUR: Ne güzel ya... Çok özendim... Ben de ailemi çok özledim. Ocak ayında arkadaşlarla tatile gittiğimiz için evde fazla kalamadım. Mayıs ayında gitmek istiyorum ama o zaman sınavlar var. Durum böyle olunca ben de finallerden sonra giderim diye Haziran'a bilet aldım. Kısa bir tatil yaparım. Zaten Temmuz'da stajım var, yine bütün yaz buradayım. Neyse, Zeynep aslında ben sana bir şey soracaktım. Sosyoloji ödevini yaptın mı ya?

ZEYNEP: Ya evet, düğüne gitmeden önce yetiştirdim ama çok hızlı yazdım, biraz aceleye geldi. Umarım hoca beğenir. Sen yazdın mı peki?

UĞUR: Sen en azından yazmışsın. Benim sosyoloji ödevi tamamen aklımdan çıkmış. Bu sabah derse gitmeden önce aklıma geldi. Aslında hiç yapasım yok ama yapmam lazım yoksa dersten kalacağım. Bugün ödev için hocadan ek süre istemeyi düşünüyorum. Sence isteyeyim mi?

ZEYNEP: Ya geçen dönem birine bir kere ek süre vermişti ama genelde vermiyor diyorlar. Bilmiyorum ki... Ruh haline bağlı biliyorsun. Eğer kızgın görünüyorsa bir şey söyleme sakın. Ama neşesi yerindeyse konuş bence. Verirse iyi olur. Ben de kendi ödevimin üstünden geçerim böylelikle.

UĞUR: İnşallah olur Zeynep ya. Konuşuruz yine, hadi hoşça kal.

ZEYNEP: Hoşça kal!

B2 LEVEL LISTENING TEXT FOR THE SECOND PILOTING

SPIKER: “Bir Varmış, Bir Yokmuş” Masal Şenliği’nde geri sayım başladı. Beşiktaş Belediyesi ve Çocuk Masalları Akademisi’nin 4-5 Haziran tarihlerinde Akatlar Sanatçılar Parkı’nda düzenlediği Masal Şenliği’ni, etkinlik koordinatörü Ayşegül Dede’yle konuşacağız. “Bir Varmış, Bir Yokmuş” Türkiye’nin ilk masal şenliği, nasıl oluştu bu proje, nereden çıktı?

AYŞEGÜL DEDE: Evet, sizin de dediğiniz gibi Masal Şenliği Türkiye’de ilk defa yapılan bir etkinlik ve bizler de bu organizasyonun bir parçası olmaktan gurur duyuyoruz. Çok emek verdiğimiz bir iş oldu. Masal Şenliği bizim 3 senedir üzerinde çalıştığımız bir proje. Projenin organizasyonunu Beşiktaş Belediyesi ile birlikte yaptık. Bu projenin çıkış noktası ise şöyle. Masallar etkili bir iletişim aracı, ve şu anda çok popülerler. Bir sürü masal son zamanlarda pek çok sinema filmine ve televizyon dizisine çevrildi ve büyük ilgi gördü. Bu da bize gösterdi ki masallar hem çocuklar için hem de yetişkinler için çok büyüleyici, çok çekici... Biz de Masal Şenliği’yle masal anlatımı sanatına katkıda bulunmak istedik. Masalların ve masal anlatma geleneğinin korunup yaşatılması ve bunların gelecek kuşaklara aktarılmasına çok önem veriyoruz. Dolayısıyla Çocuk Masalları Akademisi olarak çok heyecanlıyız, şenliği dört gözle bekliyoruz.

SPIKER: Biz de dört gözle bekliyoruz. Peki şenliğe dair nasıl tepkiler almayı umuyorsunuz?

AYŞEGÜL DEDE: Öncelikle söylemeliyim ki biz kendimize güveniyoruz. Şenlik bu sene Sanatçılar Parkı’nda çok güzel bir katılımcı ekiple gerçekleşecek. Beşiktaş Belediyesi uluslararası bir organizasyon yapmak istedi, ancak biz şimdilik sadece Türkiye’den katılımcılar davet ettik. Birazdan daha detaylı da bahsedeceğim, katılımcı ekibimize çok güveniyorum... Sanatçılar Parkı’nın da bu şenliğe ayrı bir

güzellik, masalsı bir atmosfer katacağını düşünüyorum. Hem etkinliklerimiz sayesinde hem de katılımcılarımız sayesinde şenliğin bayağı gündeme oturacağını, çok konuşulacağını düşünüyoruz. Eminim halkımızın her kesiminden ve her yaştan insanın bayağı dikkatini çekecek bir şenlik olacak.

SPIKER: Peki Masal Şenliği'nde neler oluyor, sadece masal mı anlatılıyor?

AYŞEGÜL DEDE: Hayır, başka etkinlikler de var ama masal anlatımlarımız da çok çeşitli olacak. Şöyle ki, masal anlatım performanslarının gerçekleşeceği bir açık sahnemiz var. Orada çok farklı masalcılar yer alacak. Mesela, 7-14 yaş grubu için müzikli masallar olacak. Bunun yanı sıra konuşmadan sadece beden hareketleri kullanarak, pandomimle masal anlatımları yer alacak. Çocuklar sanatçılarla birlikte pandomim yapacaklar. Yine resimli masal anlatımları göreceğiz. Ayrıca doğaçlama masal etkinliklerimiz olacak. Yetişkin seyircileri de anlatıma dahil ettiğimiz, katılımcıların önceden hazırlanmadan, içlerinden geldiği gibi oynayıp anlattıkları masallar olacak. Çocuklar anne-babalarıyla doğaçlama masal anlatımına katılabilecekler. Gelecek yıl belki farklı bir şey yapıp yetişkinlerin ve çocukların da katılacağı dansla masal anlatımı da yapılabilir, ama henüz bilmiyoruz. Bunların haricinde sahnenin etrafında standlarımız olacak. Bu standlara da müzeler katılacaklar. Biz şenliğimizi mümkün olduğunca kapsamlı, gelen kişilerin çok yönlü eğlenip öğrenebildikleri bir şenlik haline getirmek istiyoruz ve bunun en güzel yollarından birinin müzelerin katılımı olduğunu düşünüyoruz. Çünkü müzeler sadece tarihi eserlerin görüldüğü yerler değil. Özellikle yurtdışında müzelerin sanat eğitimin sık sık bir parçası olduğunu görüyoruz ve bir sanat dalının çocuklarla tanıştırıldığı, bulunduğu mekanlar müzeler aslında. Masal anlatıcılığı da kültürümüzde çok önemli yeri olan değerli bir sanat. Dede Korkut masallarını, Binbir Gece masallarını düşünün... Biz de bundan yola çıkarak bu güzel sanatı çocuklarımıza yine müzelerin

atölyeleriyle birleştirerek sunacağız şenlikte. Ayrıca bir de kapalı çadır alanımız olacak, orada da 7 artı yaş için, yani 7-14 yaş için atölye çalışmaları olacak.

SPIKER: 4-5 Haziran'da Akatlar Sanatçılar Parkı'nda dediniz ve etkinliklerden biraz bahsettiniz ama, Masal Şenliği'nin programını da merak ediyoruz. Kimler ne zaman, neler anlatıyor?

AYŞEGÜL DEDE: Kimler ne anlatıyor? Bir kere sabah 12'den 6'ya kadar çocuklara yönelik masallar anlatıyoruz. Bu masallar arasında, dünya klasikleri var, çağdaş masallar var, Anadolu masalları var... Anadolu'dan gelen çok önemli masalcıları ağırlayacağız. Bu işin duayenleri olan, uluslararası alanda ün yapmış, masal anlatıcılığı sanatını günümüze taşıyan ustalarımız var. Ayrıca saat 6'dan sonra da yetişkinlere yönelik masallar olacak. Korku masalları olacak. Çok eğlenceli, güzel başka etkinlikler olacak. Ayrıca Demet Tuncer'in sahnesiyle çok keyifli bir gece geçireceğiz. Akşam 10'a kadar biz Sanatçılar Parkı'nda olacağız.

SPIKER: Çok güzel... Ben bir de şenlik geçidinin olacağını duydum.

AYŞEGÜL DEDE: Evet, doğru duymuşsunuz. Cumartesi günü harika, masal gibi bir kortej geçidimiz olacak. Bu kortej geçidiyle birlikte sabah saat 11'de şenlik açılışımızı yapmış olacağız. Umuyorum ki büyük bir kalabalıkla çok eğlenceli bir yürüyüş olacak. Başka şehirlerden gelen farklı misafirlerimiz de olacak. Örneğin Eskişehir Masal Şatosu ekibiyle birlikte bize eşlik edecek. Aynı zamanda diğer müzeler de ekipleriyle o geçide katılacaklar. İstanbul'dan Pera Müzesi, Oyuncak Müzesi, Ankara'dan Somut Olmayan Kültürel Miras Müzesi gelecek yine bize eşlik etmek için... Ve Çocuk Masalları Akademisi ekibi olarak biz de ev sahipliği yapacağız. Şenlik geçidi belediye başkanımızın konuşmasıyla sona erecek. Hemen arkasından masal performansları başlayacak. Gündüz çocuklara akşamsa yetişkinlere yönelik tamamen ücretsiz, bilet almaya gerek olmayan etkinliğimize herkesi

bekliyoruz.

SPIKER: Anlattıklarınız gerçekten çok heyecan verici ve şenlik çok eğlenceli geçeceęe benziyor. Peki bu şenlięi düzenlemenizde daha farklı amaçlar var mıydı?

Yani toplumsal anlamda düşündüğünüz şeyler var mıydı?

AYŞEGÜL DEDE: Eee, tabi ki vardı...Biz aynı zamanda bir sivil toplum kuruluşuyuz, biliyorsunuz... Açıkçası gözlemlediğimiz kadarıyla, yaratıcı etkinliklere Türk toplumunun bütün kesimleri ilgi göstermiyor. Biz Türk halkı olarak bir araya gelip açık alanlarda tanımadığımız insanlarla birlikte kaygısızca eğlenmeye pek alışık değiliz. Çok farklı topluluklara ve her çeşit insana açık bu şenlikte halkımızın çeşitli kesimlerini bir araya getireceğimizi düşünüyoruz. Sosyal kutuplaşmanın özellikle arttığı çağdaş dünyada bu tür etkinliklerin toplulukları kaynaştırıcı gücü olduğuna inanıyoruz.

SPIKER: NTV Radyo kültür-sanat köşesine katkınız için teşekkürler. Görüşmek üzere.

AYŞEGÜL DEDE: Ben teşekkür ederim, görüşmek üzere.

APPENDIX D

TEST SPECIFICATIONS FOR THE SECOND ADMINISTRATION

TEST SPECIFICATIONS FOR A1 LEVEL TASK

TASK CHARACTERISTICS	
Task description	Instructions are provided both in oral and written form about the task and the responses; the number of dialogues, the time allocated before listening to read the questions, the format of responses and the number of times the dialogues will be played. The dialogues are contextualized with the help of titles indicating the location of the dialogues. Questions are given on paper and presented before the listening starts. Students are given 1 minute to read the questions before the listening. They write the answers in the blanks provided on the answer sheet. They have 2 minutes to check their answers after the listening finishes.
Skill focus	Listening to short dialogues
Related TLU task	Comprehending dialogues between persons with whom students are likely to have interaction during their academic studies (e.g. professors, academic advisors, fellow students, shopkeepers near school, university clerks, university doctors, etc.)
Task type	Listening to short dialogues and writing short answers to the open-ended questions. Dialogue is heard only once.
Instructions to candidates	Now you are going to listen to 3 short dialogues. You are going to listen to them once. Firstly read the questions in 1 minute. Write the correct answers in the blanks.
TEXT CHARACTERISTICS	
Text source	The test provider
Discourse purpose	Informative and exploratory
Domain	Public & Educational
Discourse type	Short dialogues
Text input appears to be genuine	Yes
Content / subject knowledge	General / School environment
Cultural specificity	Neutral
Nature of information	Only concrete
Channel of presentation	Aural (recorded text) and visual (questions on paper)
Text speed	Slow

Text length	Short
Grammar	Simple sentences only
Vocabulary	Frequently used simple vocabulary items related to school and school environment
Number of participants	2 per dialogue
Accent standard	Standard
Language of input	Turkish
Clarity of articulation	Clear
How often played	Once
Comprehensible by learner at CEFR level	A1
ITEM CHARACTERISTICS	
Estimated CEFR level of items	A1
Item type	Short-answer questions
Number of items	6
Response format	Single-word short answers in the blanks provided on the answer sheet
Scoring parameters	Objectively scored dichotomous items (0 or 1) with each item equally weighted. Small spelling mistakes are ignored as long as they do not interfere with meaning or unless the answer is a proper name or a very high frequency word.
Targeted listening skills	
Items 1-6	Listening for specific factual information clearly stated

TEST SPECIFICATIONS FOR A2 LEVEL TASK

TASK CHARACTERISTICS	
Task description	Instructions are provided both in oral and written form about the task and the responses; the type of listening text, the time allocated before listening to read the questions, the format of responses and the number of times the dialogue will be played. The dialogue is contextualized with the help of a title indicating the topic of conversation. Item stems and options are given on paper and presented before the listening starts. Students are given 1 minute to read the questions before the listening. They choose the correct answer from the options provided on the answer sheet or write the correct answers in the blanks. They have 2 minutes to check their answers after the listening finishes.
Skill focus	Listening to a dialogue between two classmates
Related TLU task	Comprehending important information in long dialogues between students and persons with whom students are likely to have interaction during their academic studies and around school environment.
Task type	Listening to a dialogue and choosing the correct answer from the options given or writing the correct answers in the blanks. Dialogue is heard only once.
Instructions to candidates	Now you are going to listen to a dialogue between two friends at a dormitory. You are going to listen to it once. Firstly read the questions in 1 minute. Choose the correct answer or write the correct answers in the blanks.
TEXT CHARACTERISTICS	
Text source	The test provider
Discourse purpose	Informative, exploratory and expressive (of individual)
Domain	Public & Educational
Discourse type	A short/moderate length dialogue
Text input appears to be genuine	Yes
Content / subject knowledge	General / School environment
Cultural specificity	Neutral
Nature of information	Only concrete
Channel of presentation	Aural (recorded text) and visual (questions on paper)

Text speed	Slow to moderate
Text length	Short/moderate
Grammar	Mostly simple sentences with a few co-ordinate clauses
Vocabulary	Frequently used simple and average difficulty vocabulary items related to school and courses
Number of participants	2
Accent standard	Standard
Language of input	Turkish
Clarity of articulation	Clear
How often played	Once
Comprehensible by learner at CEFR level	A2
ITEM CHARACTERISTICS	
Estimated CEFR level of items	A2
Item type	Three-option multiple choice, gap-filling and short-answer questions
Number of items	8
Response format	Options & generally single-word short answers in the blanks provided on the answer sheet
Scoring parameters	Objectively scored dichotomous items (0 or 1) with each item equally weighted. For short answers small spelling mistakes are ignored as long as they do not interfere with meaning or unless the answer is a proper name or a very high frequency word
Targeted listening skills	
Items 1-8	• Listening for specific factual information clearly stated • Listening for main idea(s) or important information: and distinguishing that from supporting detail, or examples

TEST SPECIFICATIONS FOR B1 LEVEL TASK

TASK CHARACTERISTICS	
Task description	Instructions are provided both in oral and written form about the task and the responses; the type of listening text, the time allocated before listening to read the questions, the format of responses and the number of times the dialogue will be played. The dialogue is contextualized with the help of instructions explaining the context and a title. Item stems and options are given on paper and presented before the listening starts. Students are given 1 minute to read the questions before the listening. They choose the correct answer from the options provided on the answer sheet. They have 2 minutes to check their answers after the listening finishes.
Skill focus	Listening to a dialogue between two classmates
Related TLU task	Comprehending important information in long dialogues between students and persons with whom students are likely to have interaction during their academic studies and around school environment.
Task type	Listening to a dialogue and choosing the correct answer from the options given. Dialogue is heard only once.
Instructions to candidates	Now you are going to listen to a dialogue between two friends about their weekend activities. You are going to listen to it once. Firstly read the questions in 1 minute. Choose the correct answers.
TEXT CHARACTERISTICS	
Text source	The test provider
Discourse purpose	Informative, exploratory and expressive (of individual)
Domain	Public & Educational
Discourse type	A moderately long dialogue
Text input appears to be genuine	Yes
Content / subject knowledge	General / School environment
Cultural specificity	Neutral
Nature of information	Mostly concrete
Channel of presentation	Aural (recorded text) and visual (questions on paper)
Text speed	Medium speed
Text length	Moderately long
Grammar	A combination of simple and complex sentences with cohesive devices and linkers

Vocabulary	Average difficulty vocabulary items about daily activities and school environment, and a few less frequent words/phrases
Number of participants	2
Accent standard	Standard
Language of input	Turkish
Clarity of articulation	Clear
How often played	Once
Comprehensible by learner at CEFR level	B1
ITEM CHARACTERISTICS	
Estimated CEFR level of items	B1
Item type	Multiple-choice questions with four short options
Number of items	6
Response format	Multiple choice
Scoring parameters	Objectively scored dichotomous items (0 or 1) with each item equally weighted.
Targeted listening skills	
Items 1-6	• Listening for specifics, including recall of important details • Listening for main idea(s) or important information: and distinguishing that from supporting detail, or examples • Understanding discourse markers • Identifying and reconstructing topics and coherent structure from ongoing discourse involving two or more speakers • Determining a speaker's attitude or intention towards a listener or a topic • Making inferences and deductions at local levels

TEST SPECIFICATIONS FOR B2 LEVEL TASK

TASK CHARACTERISTICS	
Task description	Instructions are provided both in oral and written form about the task and the responses; the type of listening text, the time allocated before listening to read the questions, the format of responses and the number of times the lecture will be played. The lecture is contextualized with the help of instructions explaining the context and a title. Items are given on paper and presented before the listening starts. Students are given 3 minutes to read the questions before the listening. They choose the correct answer from the options provided on the answer sheet. They have 2 minutes to check their answers after the listening finishes.
Skill focus	Listening to an interview (radio program) between an interviewer and an interviewee
Related TLU task	Comprehending important information in long stretches of speech such as an interview which students are likely to hear in the outside world
Task type	Listening to an interview and choosing the correct answer from the options given. Dialogue is heard only once.
Instructions to candidates	Now you are going to listen to a radio program about a story festival called ‘Bir Varmış, Bir Yokmuş’. You are going to listen to it once. Firstly read the questions in 3 minute. Choose the correct answers.
TEXT CHARACTERISTICS	
Text source	Semi-scripted material
Discourse purpose	Informative, exploratory and expressive (of individual)
Domain	Public
Discourse type	A long dialogue
Text input appears to be genuine	Yes
Content / subject knowledge	Information about a festival organization
Cultural specificity	Neutral
Nature of information	Both concrete or abstract
Channel of presentation	Aural (recorded text) and visual (questions on paper)
Text speed	Normal/Fast
Text length	Long
Grammar	Mostly high-level structures with complex and co-ordinate clauses, and cohesive devices and

	linkers
Vocabulary	Both high and low frequency words about any subject
Number of participants	2
Accent standard	Standard
Language of input	Turkish
Clarity of articulation	Clear
How often played	Once
Comprehensible by learner at CEFR level	B2
ITEM CHARACTERISTICS	
Estimated CEFR level of items	B2
Item type	Multiple-choice questions with four short options
Number of items	8
Response format	Multiple choice
Scoring parameters	Objectively scored dichotomous items (0 or 1) with each item equally weighted.
Targeted listening skills	
Items 1-8	• Listening for specifics, including recall of important details • Listening for main idea(s) or important information: and distinguishing that from supporting detail, or examples • Identifying role of discourse markers in signaling structure of a text (conjunctions, adverbs, etc.) • Identifying and reconstructing topics and coherent structure from ongoing discourse involving two or more speakers • Determining a speaker's attitude or intention towards a listener or a topic • Making inferences and deductions at both local and global levels

APPENDIX E

TASK EVALUATION QUESTIONNAIRES

A1 LEVEL TASK EVALUATION QUESTIONNAIRE

1. Which skill(s) did you use while answering each question? Put a tick in the relevant box.
You can choose more than one skill for each question.

In order to answer this question correctly I had to...	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1. understand specific bits of information in the dialogue						
2. understand just the main idea(s)						
3. understand the details used to explain the main idea(s)						
4. differentiate between important and less important information						
5. understand what the dialogue is about briefly						
6. understand how information in the whole dialogue fits together						
7. pay attention to the speakers' attitude and tone						
8. understand what the speaker's intention is when using a certain sentence						
9. rely on my general world knowledge						

2. Please indicate the difficulty level of each question. Put a tick in the relevant box.

Question no	Too easy	Moderate	Difficult	Too difficult
1				
2				
3				
4				
5				
6				

3. Please state your opinion about the listening test. Put a tick in the relevant box.

1. The instructions were clear.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
2. The recording was audible.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
3. It was enough to listen to the text once.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
4. The recording was comprehensible.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
5. The text was relevant to what I listen to in real life.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
6. Please evaluate the speed of the recording.	Slow	Normal	Fast		
7. Please evaluate the difficulty level of the task.	Too easy	Moderate	Difficult	Too difficult	
8. Please evaluate the difficulty of the listening text.	Too easy	Moderate	Difficult	Too difficult	

A2 LEVEL TASK EVALUATION QUESTIONNAIRE

1. Which skill(s) did you use while answering each question? Put a tick in the relevant box.
You can choose more than one skill for each question.

In order to answer this question correctly I had to...	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
1. understand specific bits of information in the dialogue								
2. understand just the main idea(s)								
3. understand the details used to explain the main idea(s)								
4. differentiate between important and less important information								
5. understand what the dialogue is about briefly								
6. understand how information in the whole dialogue fits together								
7. pay attention to the speakers' attitude and tone								
8. understand what the speaker's intention is when using a certain sentence								
9. rely on my general world knowledge								

2. Please indicate the difficulty level of each question. Put a tick in the relevant box.

Question no	Too easy	Moderate	Difficult	Too difficult
1				
2				
3				
4				
5				
6				
7				
8				

3. Please state your opinion about the listening test. Put a tick in the relevant box.

1. The instructions were clear.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
2. The recording was audible.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
3. It was enough to listen to the text once.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
4. The recording was comprehensible.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
5. The text was relevant to what I listen to in real life.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
6. Please evaluate the speed of the recording.	Slow	Normal	Fast		
7. Please evaluate the difficulty level of the task.	Too easy	Moderate	Difficult	Too difficult	
8. Please evaluate the difficulty of the listening text.	Too easy	Moderate	Difficult	Too difficult	

B1 LEVEL TASK EVALUATION QUESTIONNAIRE

1. Which skill(s) did you use while answering each question? Put a tick in the relevant box.
You can choose more than one skill for each question.

In order to answer this question correctly I had to...	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1. understand specific bits of information in the dialogue						
2. understand just the main idea(s)						
3. understand the details used to explain the main idea(s)						
4. differentiate between important and less important information						
5. understand what the dialogue is about briefly						
6. understand how information in the whole dialogue fits together						
7. pay attention to the speakers' attitude and tone.						
8. make an inference based on the information in the text						
9. understand relations between the speakers and the situation they are in						
10. understand what the speaker's intention is when using a certain sentence						
11. understand what an unknown word/phrase means based on the information in the text						
12. rely on my general world knowledge.						

2. Please indicate the difficulty level of each question. Put a tick in the relevant box.

Question no	Too easy	Moderate	Difficult	Too difficult
1				
2				
3				
4				
5				
6				

3. Please state your opinion about the listening test. Put a tick in the relevant box.

1. The instructions were clear.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
2. The recording was audible.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
3. It was enough to listen to the text once.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
4. The recording was comprehensible.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
5. The text was relevant to what I listen to in real life.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
6. Please evaluate the speed of the recording.	Slow	Normal	Fast		
7. Please evaluate the difficulty level of the task.	Too easy	Moderate	Difficult	Too difficult	
8. Please evaluate the difficulty of the listening text.	Too easy	Moderate	Difficult	Too difficult	

B2 LEVEL TASK EVALUATION QUESTIONNAIRE

1. Which skill(s) did you use while answering each question? Put a tick in the relevant box. You can choose more than one skill for each question.

In order to answer this question correctly I had to...	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
1. understand specific bits of information in the dialogue								
2. understand just the main idea(s)								
3. understand the details used to explain the main idea(s)								
4. differentiate between important and less important information								
5. understand what the dialogue is about briefly								
6. pay attention to the speakers' attitude and tone								
7. understand how information in the whole dialogue fits together								
8. understand how certain parts are linked to others in the dialogue								
9. make an inference based on the information in the text								
10. understand what the speaker's intention is when using a certain sentence								
11. understand what an unknown word/phrase means based on the information in the text								
12. rely on my general world knowledge.								

2. Please indicate the difficulty level of each question. Put a tick in the relevant box.

Question no	Too easy	Moderate	Difficult	Too difficult
1				
2				
3				
4				
5				
6				
7				
8				

3. Please state your opinion about the listening test. Put a tick in the relevant box.

1. The instructions were clear.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
2. The recording was audible.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
3. It was enough to listen to the text once.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
4. The recording was comprehensible.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
5. The text was relevant to what I listen to in real life.	Definitely agree	Agree	Neutral	Disagree	Definitely disagree
6. Please evaluate the speed of the recording.	Slow	Normal	Fast		
7. Please evaluate the difficulty level of the task.	Too easy	Moderate	Difficult	Too difficult	
8. Please evaluate the difficulty of the listening text.	Too easy	Moderate	Difficult	Too difficult	

APPENDIX F

CONSENT FORM

Institution: Boğaziçi University

Research Title: Development of a Listening Test for Second Language Learners of Turkish

Project Coordinator: Assist. Prof. Aylin ÜNALDI

E-mail address: aunaldi@boun.edu.tr

Phone number: 0212 359 46 09

Researcher's Name: Emel TOZLU

E-mail address: emel.hakyemez@gmail.com

Phone number: 0539 859 60 29

Topic of the Project: This listening test is being prepared to assess listening skills of future learners of Turkish as a second language. This study aims to provide evidence for the validity and reliability claims of this test and the test scores. Your input will be most valuable and appreciated. If you are willing, please take the test tasks and complete the evaluation sheets right after you take them. This will approximately take an hour. Further information will be given to you by teaching assistants.

Consent: Any information from this study will be used for research purposes only and will be kept confidential. In reports to be published on this study, no information that would make it possible to identify you will be included. This is not a required component in your program, and you as learners will not be evaluated on this test. You can withdraw from this study at any time. If you agree to participate in this research please read and sign this form indicating your willingness (or not) to participate in this research and if you are willing, complete the following questions, too. If you have any questions, you can ask them any time to the project coordinator or the researcher. You can also consult the university's Ethics Committee regarding your rights in this study.

When the test results are ready, the results will be made available to all who are interested. If you have any further information, you can contact:
emel.hakyemez@gmail.com

I have read and understood the information provided above. I agree to participate in this study.

Signature: _____

Name, Surname: _____

Your Turkish Class: _____

Gender: _____

Mother tongue: _____

Country: _____

Age & place of first exposure to Turkish: _____

How long have you been learning Turkish?: _____

How often do you use Turkish outside the course?: _____

APPENDIX G

CEFR VOCABULARY DESCRIPTORS

CEFR level	Overall proficiency	Linguistic range
A1	- Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type.	- Has a very basic range of simple expressions about personal details and needs of a concrete type. - Has a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations.
A2	- Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment).	- Can control a narrow repertoire dealing with concrete everyday needs. - Has a repertoire of basic language which enables him/her to deal with everyday situations with predictable content - Has a limited repertoire of short memorised phrases covering predictable survival situations
B1	- Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.	- Has a sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and films. - Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events, but lexical limitations cause repetition and even difficulty with formulation at times. - Shows a good control of elementary vocabulary but major errors still occur when expressing more complex thoughts or handling unfamiliar topics and situations.
B2	- Can understand the main ideas	- Has sufficient range of language to

of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation.	be able to give clear descriptions, express viewpoints and develop arguments without much conspicuous searching for words, using some complex sentence forms to do so. - Can express him/herself clearly and without much sign of having to restrict what he/she wants to say. - Has a good range of vocabulary for matters connected to his/her field and most general topics.
--	---

Figure G1. CEFR descriptors for vocabulary knowledge

APPENDIX H

FUNCTIONAL DIMENSIONS OF THE LISTENING TEXTS

Table H1. Functional Dimensions of the Listening Texts

CEFR level	Functions	Examples
A1	Imparting and seeking factual information	
	Reporting (describing and narrating) declarative sentences	<i>Yurtta bir kaza geçirdim. Kapı hızlı bir şekilde koluma ve başıma çarptı.</i>
	Asking for information	<i>Sunum dosyalarının tanesi ne kadar?</i>
	Answering questions for confirmation	<i>(Merhaba, öğrenci işleri ofisi burası mı?) Evet burası.</i>
	Answering questions for information	<i>(Borcum ne kadar?) Hepsi 7 lira 75 kuruş.</i>
	Expressing and finding out attitudes	
	Expressing obligation to do something	<i>Önce internetten başvuru yapmalısınız.</i>
	Factual modality: inquiring about ability and inability	<i>Kolunuzu hareket ettirebiliyor musunuz?</i>
	Volitional: expressing want, desire	<i>Ben dosya almak istiyorum.</i>
	Emotional: expressing gratitude	<i>Teşekkürler.</i>
	Deciding on and managing courses of action: suasion	
	Requesting others to do something	<i>Onları da ekleyebilir misiniz?</i>
	Offering assistance	<i>Nasıl yardımcı olabilirim?</i>
	Socialising	
	Attracting attention & greeting people	<i>Merhaba!</i>
	Taking leave	<i>İyi günler.</i>
	Structuring discourse	
	Opening a conversation	<i>Merhaba, öğrenci işleri ofisi burası mı?</i>
	Closing	<i>Kolay gelsin, iyi günler</i>

A2	Imparting and seeking factual information	
	- Reporting (describing and narrating) declarative sentences - Asking for information - Answering questions for confirmation - Answering questions for information	<i>Sabah midem ağrıdı, doktora gittim. Gelemedim.</i> <i>Dersler nerede yapılacakmış?</i> (<i>Kitapla ilgili bir şey dedi mi hoca?</i>) <i>Evet.</i> (<i>Peki sunum kaç dakika olacak?</i>) <i>15-20 dakika sürecekmış.</i>
	Expressing and finding out attitudes	
	- Factual: agreement etc.: expressing agreement with a statement - Factual: modality: expressing how (un)certain one is of something - Volitional: expressing intention - Emotional: expressing pleasure, liking - Emotional: expressing gratitude	(<i>Sen dersleri hiç kaçırmazsın.</i>) <i>Evet haklısın.</i> <i>Evet, bence de güzel olacak.</i> <i>Ödevimiz çok yok sanırım.</i> <i>Bir tane makale yazacağız, bir tane de sunum yapacağız.</i> (<i>Yani sınav bahar tatilinden önce olacak.</i>) <i>Aa ne güzel.</i> <i>Ya çok teşekkür ederim gerçekten çok sağ ol.</i>
	Socialising	
	- When meeting people - Addressing somebody - Taking leave	<i>Selam Ayşe, n'aber?</i> <i>İyilik senden n'aber?</i> <i>Selam Ayşe, n'aber?</i> <i>Haftaya görüşürüz o zaman. Görüşürüz!</i>
B1	Imparting and seeking factual information	
	- Reporting (describing and narrating) declarative sentences - Asking for information - Answering questions for confirmation - Answering questions for information	<i>Ocak ayında arkadaşlarla tatile gittiğimiz için evde fazla kalamadım. Mayıs ayında gitmek istiyorum ama o zaman sınavlar var.</i> <i>Hafta sonun nasıldı?</i> (<i>Sosyoloji ödevini yaptın mı ya?</i>) <i>Ya evet, düğüne gitmeden önce yetiştirdim.</i> (<i>Sergiden sonra ne yaptınız?</i>) <i>Akşam da Spor Kulübünden arkadaşlar aradı.</i>

Expressing and finding out attitudes	
• Factual: agreement etc.: expressing agreement with a statement	<i>(Hadi ya. Kötü olmuş.) Evet ya, ben de çok üzıldüm ama n'apalım.</i>
• Stating whether one knows or does not know a person, thing or fact	<i>Bilmiyorum ki... Ruh haline bağlı biliyorsun.</i>
• Stating whether one remembers or has forgotten a person, thing or fact or action	<i>Benim sosyoloji ödevi tamamen aklımdan çıkmış. Bu sabah derse gitmeden önce aklıma geldi.</i>
• Factual: modality: Expressing obligation	<i>Aslında hiç yapasım yok ama yapmam lazım.</i>
• Volitional: expressing intentions	<i>Bugün ödev için hocadan ek süre istemeyi düşünüyorum.</i>
• Emotional: expressing liking	<i>Ben çok beğendim sergiyi.</i>
• Emotional: expressing interest	<i>(Cuma sabahtan İzmir'e gittim.) Hadi ya, neden?</i>
• Emotional: expressing hope	<i>İnşallah olur Zeynep ya. Umarım hoca beğenir.</i>
Deciding on and managing courses of action: suasion	
• Advising someone to do something	<i>Baya ilginç resimler vardı, vaktin olursa kesinlikle gitmelisin</i>
• Warning others to do something or to refrain from doing something	<i>Eğer kızgın görünüyorsa bir şey söyleme sakın.</i>
Socialising	
• When meeting a friend or acquaintance	<i>Merhaba Uğur, günaydın, n'aber?</i>
• Replying to a greeting	<i>İyilik, n'olsun...</i>
• Addressing somebody	<i>İyi Zeynep, sen nasılsın?</i>
• Taking leave	<i>Konuşuruz yine, hadi hoşça kal.</i>
Structuring discourse	
• Asking someone's opinion	<i>Bugün ödev için hocadan ek süre istemeyi düşünüyorum. Sence isteyeyim mi?</i>
B2	Imparting and seeking factual information
• Stating and reporting	<i>Çok emek verdiğimiz bir iş oldu. Masal Şenliği bizim 3 senedir</i>

(describing and narrating)	<i>üzerinde çalıştığımız bir proje. Projenin organizasyonunu Beşiktaş Belediyesi ile birlikte yaptık. “Bir Varmış, Bir Yokmuş” Türkiye’nin ilk masal şenliği, nasıl oluştu bu proje, nereden çıktı? Peki Masal Şenliği’nde neler oluyor, sadece masal mı anlatılıyor? Masal Şenliği’nin programını da merak ediyoruz. (Peki şenliğe dair nasıl tepkiler almayı umuyorsunuz?) Öncelikle söylemeliyim ki biz kendimize güveniyoruz.</i>
• Asking for a piece of information • Asking for confirmation or denial • Expressing curiosity • Answering questions: giving information	
Expressing and finding out attitudes	
• Expressing agreement with a statement • Expressing knowledge of a person, thing or fact • Expressing degrees of certainty • Expressing degrees of probability • Expressing wishes/wants/desires • Expressing intentions • Expressing pleasure, happiness • Expressing hope, expectation • Expressing interest • Expressing gratitude • Reacting to an expression of gratitude	<i>(Ben bir de şenlik geçidinin olacağını duydum.) Evet, doğru duymuşsunuz.</i> <i>Dansla masal anlatımı da yapılabilir, ama henüz bilmiyoruz.</i> <i>Eminim halkımızın her kesiminden ve her yaştan insanın bayağı dikkatini çekecek bir şenlik olacak.</i> <i>Gelecek yıl belki farklı bir şey yapıp yetişkinlerin ve çocukların da katılacağı dansla masal anlatımı da yapılabilir.</i> <i>Beşiktaş Belediyesi uluslararası bir organizasyon yapmak istedi.</i> <i>Bu güzel sanatı çocuklarımıza yine müzelerin atölyeleriyle birleştirerek sunacağız şenlikte.</i> <i>Şenlik bu sene Sanatçılar Parkı’nda çok güzel bir katılımcı ekiple gerçekleşecek.</i> <i>Peki şenliğe dair nasıl tepkiler almayı umuyorsunuz?</i> <i>Çok güzel...</i> <i>NTV Radyo kültür-sanat köşesine katkınız için teşekkürler.</i> <i>Ben teşekkür ederim.</i>
Socialising	

• Taking leave	<i>Görüşmek üzere.</i>
Structuring discourse	
• Introducing a theme	<i>Masal Şenliği'ni, etkinlik koordinatörü Ayşegül Dede'yle konuşacağız.</i>
• Expressing an opinion	<i>Sanatçılar Parkı'nın da bu şenliğe ayrı bir güzellik, masalsı bir atmosfer katacağını düşünüyorum.</i>
• Exemplifying	<i>Mesela, 7-14 yaş grubu için müzikli masallar olacak. Örneğin, Eskişehir Masal Şatosu ekibiyle birlikte bize eşlik edecek.</i>
• Changing the theme	<i>Ben bir de şenlik geçidinin olacağını duydum.</i>

APPENDIX I

ORDER OF THE MEAN SCORES IN THE FIRST AND SECOND PILOT ADMINISTRATIONS

Table I2. Order of Item Means in the First Pilot Administration

Items	Mean Scores
B1I2	.78
B1I6	.75
B1I7	.75
B1I3	.69
A2I2	.65
C1I1	.64
A1I4	.62
A2I4	.62
B1I5	.62
B1I8	.62
A1I2	.60
B1I9	.60
A2I1	.58
A1I1	.55
C1I4	.55
B2I7	.53
A2I3	.49
A2I6	.49
B2I1	.49
A2I5	.47
B1I1	.47
B1I10	.47
C1I6	.47
A1I3	.44
C1I7	.42
A1I6	.40
C1I9	.40
B2I8	.38
C1I10	.38
A1I5	.36
B2I2	.36
C1I2	.36
B2I4	.33
C1I8	.29
B2I3	.27
B2I6	.27
B2I10	.25
B2I5	.24
B1I4	.22
B2I9	.20
C1I3	.13
C1I5	.11

Table I3. Order of Item Means for the Lower-Level Test Takers in the Second Pilot Administration

Items	Mean Scores
A2I1	1.00
A2I2	1.00
A2I4	.94
A1I1	.81
A1I4	.81
A2I7	.75
A1I3	.69
A2I3	.56
B1I3	.56
B1I4	.56
B1I5	.56
A1I6	.44
B1I2	.44
A2I5	.38
B1I1	.38
B1I6	.38
A2I6	.31
A2I8	.31
A1I5	.25
A1I2	.13

Table I4. Order of Item Means for the Higher-Level Test Takers in the Second Pilot Administration

Items	Mean Scores
A2I1	1.00
A2I2	1.00
A2I4	1.00
A2I3	.93
A2I5	.93
A2I7	.93
A2I8	.93
B1I3	.93
B1I4	.93
B1I5	.86
B2I6	.86
A2I6	.79
B1I6	.79
B2I3	.71
B1I2	.64
B2I1	.64
B2I5	.64
B2I7	.64
B2I8	.64
B1I1	.57
B2I4	.50
B2I2	.29

REFERENCES

- Aksan, Y., Aksan, M., Koltuksuz, A., Sezer, T., Mersinli, Ü., Demirhan, U. U., . . . Kurtoğlu, Ö. (2012). Construction of the Turkish National Corpus (TNC). In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012). İstanbul, Turkey. <http://www.lrec-conf.org/proceedings/lrec2012/papers.html>
- Anderson, J. R. (2000). *Cognitive psychology and its implications* (4th ed.). New York: Freeman.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.
- Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.
- Brown, G. (1990). *Listening to spoken discourse*. London: Longman.
- Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.
- Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching and assessment*. Cambridge: Cambridge University Press.
- Cutler, A., & Clifton, C. (1999). Comprehending spoken language: A blueprint of the listener. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp.123-166). Oxford: Oxford University Press.
- Douglas, D. (2010). *Understanding language testing*. London: Hodder-Arnold
- Elliott, M., & Wilson, J. (2013). Context validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening*, Studies in language testing, 35 (pp.152-241). Cambridge: Cambridge University Press.
- Field, J. (2009). *Listening in the language classroom*. Cambridge: Cambridge University Press.
- Field, J. (2013). Cognitive Validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening*, Studies in language testing, 35 (pp.77-151). Cambridge: Cambridge University Press.

- Fulcher G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London & New York: Routledge.
- Geranpayeh, A. (2013). Scoring validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening*, Studies in language testing, 35 (pp.242-272). Cambridge: Cambridge University Press.
- Geranpayeh, A., & Taylor, L. (2008). Examining listening: Developments and issues in assessing second language listening. *Cambridge ESOL: Research Notes*, 32, 2-5.
- Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception & Psychophysics*, 38(4), 299-310.
- Gülle, T. (2015). Development of a speaking test for second language learners of Turkish (Unpublished MA Thesis). Boğaziçi University, İstanbul, Turkey.
- Khalifa, H., & Weir, C. J. (2009). *Examining reading: research and practice in assessing second language reading*, Studies in language testing, 29. Cambridge: UCLES/Cambridge University Press.
- Kinneavy, J. E. (1969). The basic aims of discourse. *College composition and communication*, 20(5), 297-304.
- Kurt, Y. (2015). Development of a reading test for second language learners of Turkish (Unpublished MA Thesis). Boğaziçi University, İstanbul, Turkey.
- Küçük, F. (2017). Assessing academic writing skills in Turkish as a foreign language (Unpublished MA Thesis). Boğaziçi University, İstanbul, Turkey.
- Lynch, T. (2009). *Teaching second language listening*. Oxford: Oxford University Press
- McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.
- McQueen, J. M. (2007). Eight questions about spoken-word recognition. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp.37-53). Oxford: Oxford University Press.
- Messick, S. (1993). Foundations of validity: Meaning and consequences in psychological assessment. *ETS Research Report Series*, 1993(2), i-18.
- Nation, I. S. P., & Newton, J. (2009). *Teaching ESL/EFL listening and speaking*. New York & London: Routledge.
- Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL quarterly*, 17(2), 219-240.

- Richards, J. C. (2007). Materials development and research: Towards a form-focused perspective. *Form-Focused Instruction and Teacher Education Studies in Honor of Rod Ellis*, 147-160.
- Rost, M. (2013). *Listening in language learning*. New York & London: Routledge.
- Secolsky, C., Buchanan, W., & Drane, W. (11-15 Oct, 2015). The Evolution of Validity and Modern Psychometrics: Do We Need to Revisit Item Validity? Paper presented at 41st Annual Conference (2015)
The Three Most Important Considerations in Testing: Validity, Validity, Validity. University of Kansas, Lawrence KS, United States of America.
- Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in second language acquisition*, 31(04), 577-607.
- Taylor, L. (2013). Introduction. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening*, Studies in language testing, 35 (pp.1-35). Cambridge: Cambridge University Press.
- Trim, J. L. M. (2009). *Breakthrough*. Retrieved from https://www.coe.int/t/dg4/linguistic/Source/FinalBreakthrough%20specification_6Nov01.rtf
- Van Ek, J., & Trim, J. L. M. (1991). *Threshold 1990*. Cambridge: Cambridge University Press.
- Van Ek, J., & Trim, J. L. M. (1991). *Waystage 1990*. Cambridge: Cambridge University Press.
- Van Ek, J., & Trim, J. L. M. (2001). *Vantage*. Cambridge: Cambridge University Press.
- van Zeeland, H., & Schmitt, N. (2012). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension?. *Applied Linguistics*, ams074.
- Weir, C. J. (1993). *Understanding and developing language tests*. New York & Toronto: Prentice-Hall.
- Weir, C. J. (2005). *Language testing and validation*. Hampshire: Palgrave MacMillan.