DEVELOPMENT OF A READING TEST

FOR SECOND LANGUAGE LEARNERS OF TURKISH

YAVUZ KURT

BOĞAZİÇİ UNIVERSITY

2015

DEVELOPMENT OF A READING TEST
FOR SECOND LANGUAGE LEARNERS OF TURKISH

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfilment of the requirements for the degree of

Master of Arts

in

English Language Education

by

Yavuz Kurt

Boğaziçi University

2015

DECLARATION OF ORIGINALITY

I, Yavuz Kurt, certify that

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

Signature .....

Date 19.08.2015

ABSTRACT

Development of a Reading Test for Second Language Learners of Turkish

The purpose of this study is to develop a reading test that measures the ability to read in Turkish as a second language. Based on the reading framework by Khalifa and Weir (2009), task specifications were developed and reading tasks with different intended proficiency levels were developed based on the task specifications. The tasks were tested both on native speakers of Turkish and on learners of Turkish from 21 different language backgrounds enrolled in intermediate and advanced level Turkish classes at Boğaziçi University. Test taker data were used to assess item characteristics, reliability and validity of the tasks. Expert judgment was employed to evaluate the reading skills measured by task items. The findings from these investigations provided preliminary evidence for the reliability and validity of the reading tasks under scrutiny.

ÖZET

Türkçeyi İkinci Dil Olarak Öğrenenler için Okuma Testi Geliştirilmesi

Bu çalışmanın amacı ikinci dil olarak Türkçe okuma becerisini ölçen bir okuma testi geliştirmektir. Khalifa ve Weir (2009) tarafından önerilen okuma modeline dayanarak, ödev tanımlamaları oluşturulmuş ve bu tanımlamalara dayanarak farklı seviyeleri amaçlayan okuma ödevleri geliştirilmiştir. Ödevler, hem Türkçeyi anadil olarak konuşanlar üzerinde hem de 21 farklı dil kökeninden gelen ve Boğaziçi Üniversitesi'nde orta ve ileri seviye Türkçe dersleri almakta olan öğrenciler üzerinde denenmiştir. Sınava girenlerden elde edilen veriler, okuma ödevlerini madde özellikleri, güvenirlik ve geçerlik açısından değerlendirmek için kullanılmıştır. Soruların ölçtüğü okuma becerilerini değerlendirmek için uzman görüşü alınmıştır. Bu araştırmalardan elde edilen bulgular incelenen okuma ödevlerinin güvenirliğine ve geçerliğine dair ön kanıt sağlamaktadır.

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor Assoc. Prof. Gülcan Erçetin for her support and guidance. I also want to express my deep gratitude to Assist. Prof. Aylin Ünaldı for her guidance and invaluable suggestions. Finally, I am grateful to my wife Zeynep Deniz Kurt for her encouragement, and to my friend Talip Gülle for his help.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1
1.1 Aims of the study 1
1.2 Overview of methodology 1
1.3 Significance of the study 2
1.4 Research questions 3
1.5 Overview of thesis 4
CHAPTER 2: LITERATURE REVIEW 6
2.1 Introduction 6
2.2 Validity 6
2.3 Reliability 11
2.4 Test development 15
2.5 Reading construct and factors involved in cognitive process 18
2.6 Theories of reading 24
2.7 A reading model by Khalifa and Weir 31
2.8 Conclusion 40
CHAPTER 3: METHODOLOGY 42
3.1 Introduction 42
3.2 Participants 42
3.3 Instruments 44
3.4 Procedure 47
3.5 Data analysis 48
CHAPTER 4: RESULTS 52
4.1 Content validity 52
4.2 Differences between proficiency groups 63
4.3 Item analysis and distractor efficiency analysis 64
4.4 Correlations of reading scores with other skills 71
CHAPTER 5: DISCUSSION 73
CHAPTER 6: CONCLUSION 85
6.1 Limitations 86
6.2 Suggestions for future researchers 87
APPENDIX A: LEARNER PROFILE FORM 89
APPENDIX B: READING TASKS BEFORE REVISION 92
APPENDIX C: ITEM SPECIFICATIONS 101
APPENDIX D: READING TEST EXPERT EVALUATION FORM 106
APPENDIX E: READING TASKS AFTER REVISION 108
REFERENCES 116

LIST OF TABLES

Table 1. Facets of Validity. 9
Table 2. Number of Participants Coming from Each Country. 43
Table 3. Text Topics. 44
Table 4. Number of Participants and Duration of Reading Tasks. 47
Table 5. Text Characteristics. 53
Table 6. Experts' Ratings Regarding Instructions, Questions and Texts. 54
Table 7. List of Skills. 56
Table 8. Experts' Ratings Related to Task 1. 57
Table 9. Experts' Ratings Related to Task 2. 57
Table 10. Experts' Ratings Related to Task 3. 58
Table 11. Experts' Ratings Related to Task 4. 59
Table 12. Experts' Ratings Related to Task 5. 60
Table 13. Number of Occurrences of Aimed and Indicated Skills. 60
Table 14. Overlap between Intended Skills and Expert Judgments (Task 1). 61
Table 15. Overlap between Intended Skills and Expert Judgments (Task 2). 61
Table 16. Overlap between Intended Skills and Expert Judgments (Task 3). 61
Table 17. Overlap between Intended Skills and Expert Judgments (Task 4). 62
Table 18. Overlap between Intended Skills and Expert Judgments (Task 5). 62
Table 19. Percentage of Overlaps. 62
Table 20. Results of Independent Samples t-tests. 63
Table 21. Summary of Score Characteristics. 65
Table 22. Item Statistics for Reading Task 1. 65
Table 23. Distractor Efficiency Analysis for Task 1. 66
Table 24. Item Statistics for Reading Task 2. 67
Table 25. Distractor Analyses for Task 2. 67
Table 26. Item Statistics for Reading Task 3. 68
Table 27. Distractor Analyses for Task 3. 69
Table 28. Item Statistics for Reading Task 4. 69
Table 29. Distractor Efficiency Analysis for Task 4. 70
Table 30. Item Statistics for Reading Task 5. 70
Table 31. Spread of Responses for Task 5. 70
Table 32. Mean Difficulty and Cronbach's Alpha Values. 71
Table 33. Relationship between Sub-tests. 72

CHAPTER 1

INTRODUCTION

The process of language test production is influenced by the theory that explains the construct or skill being assessed. Reading has been explained in different ways such as a bottom-up, top-down or interactive process. Khalifa and Weir (2009) suggest that reading is carried out either carefully or expeditiously and either at global or local level. Different kinds of reading are employed based on the purpose of the reading. The present study attempts to operationalize reading skills with reading tasks depending on the reading model by Khalifa and Weir (2009). The tasks were investigated in terms of the two main concerns of assessment: validity and reliability.

1.1 Aims of the study

The aim of this research is mainly to investigate the validity and reliability of the five reading tasks that were developed in order to assess reading proficiency in Turkish as a foreign language. The target population is foreign students who are in a Turkish university and learn Turkish for different purposes such as daily use, personal interest or receiving education in Turkey. The test is designed to assess general reading proficiency in Turkish from B1 to C2 levels on the Common European Framework of Reference (CEFR, Council of Europe, 2001) scale. Thereby, this study aims to be the first step to develop the reading component of a test of Turkish, and it is the part of a larger study which also involves listening, writing and speaking components.

1.2 Overview of methodology

The reading construct was operationalized through a process of listing reading skills that are explained by the theoretical definition of the construct, comparing these skills to different proficiency levels specified by the CEFR, and producing task

specifications to assess these skills. Based on specifications, items were developed and elaborated on under the supervision of experts. Evidence related to reliability and validity was collected through statistical procedures and qualitative techniques from multiple sources. The scores from tasks were used for statistical analyses to elicit information about internal reliability, item characteristics, criterion-related validity, and correlational relationships with the other components of the test. Moreover, expert judgment was incorporated to look for both quantitative and qualitative evidence related to content validity and task design.

1.3 Significance of the study

Although there is not any research about testing Turkish as a foreign language, a number of examinations are in practice. “Distance Turkish Test” is the most known and widespread examination that aims to measure Turkish language proficiency of foreigners. It is an internet-based test devised by Ankara University’s Turkish and Foreign Languages Research and Application Centre (TÖMER) and offers a device for adult learners of Turkish to assess and certificate their language skills. It is administered at five different levels from A1 to C1 on the CEFR scale. The test has six sections which are reading, writing, speaking, listening, interaction, and grammar. The Turkish Proficiency Exam (TPE) is another exam testing Turkish as a foreign language. It is developed by the Yunus Emre Institute Exam Center. The institute’s official website indicates that the exam aims to assess the language proficiency of individuals learning Turkish as a foreign or native language, and thereby facilitating the admission of foreign students into Turkish educational institutions. Finally, TELC, which stands for The European Language Certificates, is another organization that offers over 70 different examinations, in eleven languages including Turkish. The exam is administered at five proficiency levels from A1 to

C1. These tests are administered to learners with a variety of language, educational, and cultural backgrounds. It is not clear to what extent learner characteristics are taken into account in the development of the tests. In addition, it is not clear what theoretical framework of reading was utilized in the design of the tests except for TELC. It is indicated on TELC's website that the examination is based on a theoretical construct, e.g. on a model of communicative competence, but no further information is provided.

Significance of the present study lies in its attempt to address the shortcomings of the available tests. The proposed test is designed for those learners learning Turkish in an academic setting. They are university level students and they usually learn Turkish in the university environment. Since they use English for academic purposes day to day, they are familiar with academic tasks. Although they do not necessarily need Turkish for academic purposes, the academic background of the target population has been taken into consideration in the design of test tasks. In addition, the reading tasks in the proposed test are designed within the theoretical framework of Khalifa and Weir (2009).

1.4 Research questions

The first research question aims to investigate the content validity of the five reading tasks under scrutiny. The second question investigates whether the tasks efficiently discriminate between higher and lower level proficiency levels. The third question is related to item characteristics of the reading tasks. Finally, the fourth question investigates the relationship between the sub-tests. Specifically, the following research questions were addressed:

1. Do the experts agree on the operations measured by the test items as specified by the test writers?

2. Do the test tasks differentiate between higher and lower proficiency groups?
3. What are the psychometric characteristics of the items for each reading task?
    a. What levels of item difficulty are reflected by the scores of the test takers?
    b. To what extent do the items discriminate between test takers' reading abilities?
    c. What are the internal reliability values for each task?
4. Do scores on the reading test correlate with the scores obtained from listening, writing and speaking tests administered to the same group?

For the first research question, it is hypothesized that the raters will mostly agree on the intended skills tested by each task. For the second research question, it is expected that the higher proficiency group will outperform the lower proficiency group on each reading task. Since the third research question is an exploratory one, there is no hypothesis regarding the third question. Finally, for the fourth research question, it is hypothesized that reading scores should be positively correlated with listening, writing and speaking scores given L1 and L2 research findings pointing to the close relationship of reading with listening (Hirai, 1999; Diakidoy, Stylianou, Karefillidou, & Papageorgiou, 2005; Wise et al., 2007), writing (Ahmed, 2011; Eisterhold, 1990; Grabe, 1991; see Hirvela, 2004 for an overview) and speaking (Liao, Qu, & Morgan, 2010).

1.5 Overview of thesis

Chapter 1 is an introductory chapter. In Chapter 2, a review of literature is presented regarding validity and reliability issues, test development process, reading theories, and the reading model proposed by Khalifa and Weir (2009). Chapter 3 describes the methods used in the study in detail. Chapter 4 reports the results regarding the four research questions. It presents the information extracted through expert judgment,

describes how different proficiency groups performed on the tasks, and investigates how items function with the group of participants by presenting difficulty and discrimination indices as well as distractor efficiency of the items. Chapter 4 also explains the relationship between the sub-components of the test, i.e. the correlational relationships between reading, listening, writing and speaking components. Chapter 5 presents the discussion of the results from the analyses. Finally, Chapter 6 summarizes the study and presents concluding remarks.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents a review of the literature on fundamental issues regarding developing a reading test. First, validity and reliability issues are explored. The main reason to test a person's language ability is to interpret their score as an indicator of what they know or can do (Bachman, 2004). Therefore, we give decisions based on test scores because we believe that the score reflects the language ability of the test taker in real life. Namely, we do reasoning from students' behavior (performance) to estimate their competence (Mislevy, 1996). We also expect that the test produces consistent measures of the ability we want to assess (Bachman, 2004). This means that a test should produce similar results under different conditions in the testing procedure as long as the test taker's ability level does not change.

After validity and reliability issues, stages of producing a new test are explored with the aim of explaining a proper test development process. Then, reading construct is discussed by providing different definitions and discussing the factors involved in the cognitive processes of reading. Finally, after presenting an overview of reading theories that have been proposed since 1950s, the reading model by Khalifa and Weir (2009) is explored.

2.2 Validity

One of the concerns in the process of designing and developing a language test is how useful it is (Bachman & Palmer, 1996). Among the qualities that Bachman and Palmer (1996) argue to be related to usefulness of a language test such as reliability, authenticity, impact, practicality and interactiveness, validity stands as a very important aspect that certainly needs to be responded.

Validity is defined as the "degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of the interpretations and actions based on test scores" (Messick, 1989, p. 13). Messick (1990) explains that validity is a degree, not an existing or non-existing feature; therefore, validation is a continuing process, and validity is a "summary of both existing evidence for and the actual as well as potential consequences of score interpretation" (p. 2). Messick (1989) notes that traditionally evidence regarding validity is categorized as construct, content, and criterion related.

Construct validity is about any evidence regarding the interpretation of score meaning and it has been accepted to be a unifying concept, so all sources of validity evidence are actually a part of construct validity (Messick, 1987). Messick (1995, p. 745) explains that there are six aspects of construct validity: content, substantive process, score structure, generalizability, external relationships, and testing consequences. He explains that content validity or content aspect of construct validity refers to evidence regarding content relevance and representativeness. Substantive aspect of construct validity refers to the theoretical rationales behind observed consistencies in test responses. Score structure suggests that scoring structure should be parallel with the structure of the construct domain. Generalizability aspect refers to the consistency of scores and interpretations across different conditions. External relationships of a test are criterion related evidence. It refers to the consistency of test scores with real life performance or scores from another test. Lastly, the testing consequences aspect is about evaluating the consequences of test use and score interpretation.

There are many sources of construct validity evidence. For example, Alderson, Clapham, and Wall (1995) mention that one of the approaches is the

correspondence with theory. They explain that the main concern in this approach is whether the test successfully operationalizes the theory, which can be investigated through expert judgment. Experts can be provided with some definition of the underlying theory and asked to make judgments after examining the test. The researchers state that a second approach for construct validation is internal correlations. The correlations between the sub-parts of a test are not supposed to be high because they are supposed to measure different sub-constructs. Furthermore, the sub-parts need to be correlated with the total test in order to provide more evidence for construct validity (Alderson et al., 1995). There are studies that are in line and in conflict with these correlational assumptions. For example, Liao, Qu, and Morgan (2010) analyzed data from more than 12,000 TOEIC test takers, and found the following correlations: .76 between reading and listening, .57 between reading and speaking, and .61 between reading and writing. On the other hand, Wang (2008) analyzed the scores of 57 undergraduate students on College English Test Bad 4 (CET-4). The data suggested that students' reading skills were not related to writing ($r = .021$, $p > .05$) and listening ($r = .053$, $p > .05$). Factor analysis which examines the factors that influence the performance on a test, and multitrait-multimethod which is based on correlational procedures, are other approaches to test construct validity. Weir (2005) states that reliability can also be regarded as one form of validity evidence and he prefers using the term "scoring validity" to emphasize his point. He indicates that scoring validity comes from such parameters as difficulty, discrimination, and internal consistency.

In order to look for content related evidence, a common practice is experts comparing the test content with its specifications, teaching syllabus or curriculum (Alderson et al., 1995). Bachman (2004) emphasizes that an important kind of

evidence to support content representativeness lies in the design of the test, i.e. test specifications and the tasks that are based on these specifications. For this purpose, careful design and development procedures should be followed. However, like Alderson et al. (1995), Bachman (2004) suggested that expert judgment can also be employed to investigate content representativeness. He points out that researchers, curriculum developers, language teachers or language testers can provide such type of expertise to give opinion about the abilities that each task measures. Provision of such judgment can be either verbal or through a rating scale (Bachman, 2004). Then, comparing these ratings to the target domain to see how representative a task is can be very useful to get insights into the content validity of the task. Thus, the extent the experts' judgments comply with each other and with test specifications regarding what areas of ability are measured can be presented as the evidence for content validity. However, Bachman also indicates that such type of approach has the disadvantage of possible disagreements between the experts.

Messick's (1989) progressive matrix of validity highlights the issues to be taken into account when making any judgment about the validity of a measure. Messick's (1989) view, illustrated in Table 1, summarizes the concept of validity including test interpretation and test use on evidential and consequential bases.

Table 1. Facets of Validity.

	Test Interpretation	Test Use
Evidential Basis	Construct Validity (CV)	CV+ Relevance/Utility (R/U)
Consequential Basis	CV+ Value Implications (VI)	CV+ R/U+ VI+ Social Consequences

Source: Messick, 1989, p. 20.

Evidential basis for test interpretation refers to evidence for score meaning. Such evidence is basically obtained through observing how the scores on a measure represent the construct (Hubley & Zumbo, 2011). Messick (1995) points out that

evidence and rationales to back up the trustworthiness of score interpretation form the evidential basis, and this evidential basis is construct validity. He adds that evidence regarding the "relevance of the scores to applied purpose, and the utility of scores in the applied setting" enhance the evidence for score meaning (p. 748). Relevance of test items to the intended score interpretation should be evaluated by taking into account the whole testing procedure which includes "specification of the construct domain", "typical behaviors", and "underlying processes" (Messick, 1989, p. 38). Utility of scores refers to how useful the testing is in reflecting target domain performance (Messick, 1989).

"Consequential basis of test interpretation is the appraisal of value implications of score meaning" (Messick, 1995). Messick points out that values (that are usually not evident) are attached to construct label, the theory underlying the construct and ideologies that influence the theory. He further explains that "The value implications of score interpretation are not only part of score meaning, but a socially relevant part that often triggers score-based actions and serves to link the construct measured to questions of applied practice and social policy" (p. 748). Therefore, he suggests that theoretical implications and value implications of test interpretation should be proportional. Ideologies and value implications may change across individuals or groups, but when a dialectical approach is adopted by proposing rival perspectives, it is possible to subject constructs or theories to an empirical grounding or debate (Messick, 1989, p. 62–63).

Consequential basis of test use includes the social consequences of testing. Messick (1995) advises that one way to see potential side effects of test use is comparing alternative proposals of test use in terms of their benefits and risks. The counterproposals of a proposed test use may reveal strong and weak sides of the

intended test use. He emphasizes that adverse social consequences of test use does not directly make a test use invalid, but "that adverse social consequences should not be attributable to any source of test invalidity such as construct underrepresentation or construct-irrelevant variance" (p. 748). The reason is that such consequences influence the validity of score interpretation which is a part of construct validity.

Messick (1995, p. 742) explains construct underrepresentation and construct-irrelevant variance as two major threats to construct validity. When the construct assessment is not comprehensive enough and cannot adequately cover the target domain, the construct is not represented well in the test, which is called construct underrepresentation. Construct-irrelevant variance, on the other hand, may appear in two forms as construct-irrelevant difficulty and construct-irrelevant easiness. Messick (1995) defines the former one as task aspects that are not directly construct-related and make the task irrelevantly difficult for some groups. The latter one, on the other hand, means that the clues in task formats, not the construct in focus, enable individuals to respond correctly.

2.3 Reliability

Reporting the degree to which a test is reliable is a fundamental part of developing a new language test (Brown, 2005). Bachman (2004) defines reliability as "consistency of measures across different conditions in the measurement procedure" (p. 153). In other words, reliability is a measure of how an assessment tool consistently measures learning. Scores on a test will be influenced by a number of factors such as testing procedure, non-parallel testing conditions or non-parallel test forms. Bachman (2004) categorizes sources of score variance, which is not due to the ability being measured, as personal characteristics, test method and random factors. Variations in scores that are not due to the ability we aim to measure are thought to be measurement errors

Bachman, 2004). Therefore, ideally the only variation in scores is supposed to be a result of variation in the ability being measured. The lower the measurement error, the more reliable a test is thought to be (Bachman, 2004).

Score variance can stem from many sources such as testing environment, administration procedures, test takers, scoring procedures, and test items (Brown, 2005). Some sources of variance are systematic while some of them are random. Random sources of variance are completely unexpected and unsystematic (Bachman, 2004). For example, Bachman states that test administrators cannot do anything about a test taker's being tired or air-conditioner's stopping functioning during the examination. Bachman (2004) writes that despite the limitations, there are measurement models to estimate error variance and inform about the reliability of a measure. He explains that estimating reliability requires two types of analysis, a logical analysis to detect potential sources of measurement error and a statistical analysis to quantify reliability estimation. Brown (2005) notes that when we know the degree to which error variance affects the results, we can also predict the reliability of a test. He also proposes a set of potential sources of measurement error to be taken into account such as noise, weather, timing, scoring subjectivity, test booklet clarity, test security, examinees' motivation, etc.

A set of scores can only be reliable at the extent it reflects test takers' level of ability; therefore, for reliability concerns, the true score variance is expected to be high while error score variance is low (Bachman, 2004). Furthermore, reliability is a prerequisite for validity since a test cannot measure precisely if it does not produce consistent results; however, even when the reliability of a test is estimated to be fairly high, this does not guarantee that it measures validly (Alderson et al., 1995).

Weir (2005) sees reliability as a form of validity evidence, and prefers the term scoring validity instead of reliability.

Reliability of a test is estimated by calculating a reliability coefficient, which can take a value between 0.00 and 1.00. A reliability coefficient that is close to 1.00 is thought to produce more consistent results. It is possible to calculate reliability coefficient through many different ways. Kumar (2012) suggests that there are two basic procedures to estimate reliability: external consistency procedures and internal consistency procedures.

In external consistency procedures two sets of scores are compared. For example, comparing equivalent forms of a test or test-retest method can provide estimates of reliability (Kumar, 2012). In both strategies, two sets of scores are correlated to see to what extent they produce parallel results. Ideally, equivalent forms of a test should be given to the same group of test takers without much time interval because learning is not supposed to be a confounding variable (Kumar, 2012). Homogeneous groups can also complete parallel forms of the test at the same time so that the effects of inconsistencies over time can be controlled (Bachman, 2004). Similarly, in test-retest method, the same group of test takers receive the same test at a time interval that should not allow much time for learning, but also should not give too little time that would enable students to remember the answers (Kumar, 2012).

In tests with multiple items, inconsistencies among items can give rise to measurement error (Bachman, 2004). Bachman explains that internal consistency refers to whether the items in a test consistently measure the same content because items measuring the same content are expected to bear similar results. Internal consistency is usually explored by calculating the Cronbach's alpha which is a

statistical coefficient obtained through inter-item correlations. Furthermore, classical item analyses (difficulty and discrimination values of items) can provide feedback to increase the internal consistency reliability of a test (Bachman, 2004). Split-half is another approach to test whether items in a test measure the same ability. By splitting the scores into two halves, it is possible to have two scores for each test taker for each half of the test (Bachman, 2004). Then the correlation between the two sets of scores can estimate to what extent the items of the test measure the same skill. It is quite important to make sure that the items to be split aim the same skill or abilities, not different aspects of a more general ability (Bachman, 2004).

In measurements where raters need to give their subjective judgments about a performance, inter-rater or intra-rater reliability should be explored to see how similar and consistent scores the two raters provide or one rater provides at different times (Brown, 2005). Such a way to estimate reliability is frequently employed when measuring productive abilities such as writing and speaking.

Bachman (2004, p. 157) states that the measurement models to estimate reliability highly simplify the situation; therefore, we usually explore only a few sources of measurement error. This may lead to unintentionally ignoring important sources of error variance for a given test. He also draws attention to the fact that these sources of variance may be interacting with each other, forming a combined effect on scores. Bachman (2004) also points out that Classical Test Theory assumes that measurement error is the same across all levels of ability; however, research has shown that scores are reliable to different extents at different ability levels. Brown (2005) reminds that reliability estimates are based on a particular group of people; therefore, the estimate can only be related to that particular group or to very similar groups.

2.4 Test development

Test development involves the process of creating and using a test, which starts with initial conceptualization and design, and results in one or more archived tests and the results of their use (Bachman & Palmer, 1996). The process can be quite informal in low-stakes tests; however, it requires more work, involving extensive trialing and revision, in high-stakes tests that are planned to be used for important decisions (Bachman & Palmer, 1996). According to "Language examining and test development", a paper prepared under the direction of Milanovic (2002) for CEFR for languages, the process starts with a perceived need for a test, and planning, designing, development, operational and monitoring phases follow it.

After the initial perception that a new test is necessary, the first step is planning. Planning of a test is based on the needs for the test and the group of test takers for whom the test is intended for (Milanovic, 2002). In this phase, the purpose is to neatly analyze the potential candidates and the potential purposes of the test's use (Milanovic, 2002). Therefore, careful planning is a means for assuring that the test will be useful for its intended purpose (Bachman & Palmer, 1996). Downing (2006) states that clear test planning is a crucial step for successful preparing, administering, scoring, and analyzing a test. Hughes (2003) states that the primary step in testing is being clear about what to measure and to what purpose; therefore, a number of crucial questions need to be answered in the planning phase such as: "What kind of test is it to be?", "What is its precise purpose?", "What abilities are to be tested?", "How detailed must the results be?" (p. 59).

The abilities to be tested are determined based on the purpose of the test. Alderson (2000) states that "Every test is intended to measure one or more constructs." (p. 118). Definition of the target construct or the theoretical model of the

construct comes from a theory. The theory explains the construct, sub-constructs and the relation between them (Alderson, 2000). Especially ability and achievement tests rely on content related validity; therefore, the content domain should be carefully defined because any inadequacy at this stage cannot be compensated for (Downing, 2006).

In the design phase, initial test specifications are produced (Milanovic, 2002). "Test specifications provide the link between theoretical and operational definitions since the test specifications provide the guidance to the test writes, as well as to test users" (Alderson, 2000, p. 124). Operationalization involves developing "task specifications" and "a blueprint" indicating how the tasks will be arranged in the test (Bachman & Palmer, 1996). Alderson at. al. (1995) state that "A test's specifications provide the official statement about what the test tests and how it tests it." (p. 9). It is also stated that two forms of tests specifications can be prepared; one with information that will interest only test writers and a second one for test takers and test users. Specifications should include information, along with test purpose and target population, about how many sections it has, test content and method, text types to be employed, what language skills to be tested, what sort of tasks are required, how many items there are for each section, and what kind of rubrics are to be used (Alderson et al., 1995).

Downing (2006) suggests that test specifications and their rationales form a basis for systematic test development activities and for content validity evidence which is necessary to support score inferences regarding target knowledge domain or performance. Downing (2006) indicates that once the test specifications are produced, item development and test assembly are the next concerns. He indicates that choosing appropriate item formats, writing example initial items, and creating

test forms are carried out at this step. Alderson et al. (1995) warn that items should be based on the specifications, not the previous tests; however, test writers should consult to the previous tests, especially when producing text based items, in order to avoid overusing similar materials with similar content. They also emphasize that different item types should be tried by testing the same skills with different methods. This can enable testers to see which item types are more effective or whether employing multiple item types will bear more reliable results.

It is not possible to cover the whole target content in one form of the test; therefore, the sampled abilities should be tracked in each form in order to equally and adequately cover the content specifications across different forms (Hughes, 2003). Hughes (2003) also suggests item moderation in which at least two colleagues examine the produced items to detect weak parts.

The development phase also covers pretesting (Milanovic, 2002). Pretesting is a general term that refers to trials of the test on natives or on groups that are representative of the target population (Alderson et al., 1995). Bachman and Palmer (1996) note that the purpose of such trials is collecting information about the usefulness of the test. Based on the information from trials, the potential needs for minor editing or global revisions can be revealed. Such trials are necessary to see whether the test is working in the anticipated way (Bachman & Palmer, 1996). For example, performance of multiple choice items is especially hard to predict because presence of a variety of correct and incorrect answers leads to potential ambiguities and disagreements (Alderson et al., 1995, p. 74).

In the operational phase, the test is made available to candidates (Milanovic, 2002). Operational test use both aims to accomplish intended purpose of the test and collect more information about test usefulness (Bachman & Palmer, 1996). The

results from such administrations are, for example, used for item analysis or to investigate reliability of the test and validity of test use (Bachman & Palmer, 1996).

After a test has become operational, the monitoring phase begins (Milanovic, 2002). This phase involves monitoring the results from live administrations, collecting regular feedback from test takers and school teachers, and also involves any research to see what kind of improvements can be done on the test or its administration (Milanovic, 2002). Once the test begins to be routinely administered, all the test tasks or items should be archived so that a bank of test tasks is built in order to facilitate the development of subsequent tests (Bachman & Palmer, 1996). Downing (2006) suggests that every testing program needs to be systematically documented in technical reports that describe important aspects of test development, administration, scoring, reporting, analyses and evaluation. He explains that such documentation can provide validity evidence and identify potential threats to validity.

2.5 Reading construct and factors involved in cognitive process

In order to develop a reading test, the construct of reading needs to be understood well. Alderson (2000, p. 117) states that constructs come from a theory of reading, and definitions of reading construct in assessment may change based on the testing purpose. Grabe and Stoller (2002) suggest that reading is "the ability to understand information in a text and interpret it appropriately" (p.17). Goodman (2001) talks about reading as a dynamic and constructive process while Urquhart and Weir (1998) provides a more specific definition of reading: "the process of receiving and interpreting information encoded in language form via the medium of print" (p. 22). Grabe (2009) also sees reading as a process and probes into the construct of reading by discussing the nature of reading process, and he explains that reading is a rapid,

efficient, comprehending, interactive, strategic, flexible, purposeful, evaluative, learning, and linguistic process.

Koda (2005) notes that "comprehension is achieved through the integrative interaction of extracted text information and a reader's prior knowledge" (p. 4). The meaning constructed by readers depending on the same text will vary (Goodman, 2001) because each reader brings their own sense to a text. Based on a multi-level text representation, Kintsch and Rawson (2005) suggest that a reader constructs a literal meaning from the text, but it is not sufficient to build a deep understanding. Therefore, the explicitly stated content of the text is combined with background knowledge and purpose of the reader. This is a mental model of the situation and it is not restricted to verbal domain, frequently involving imagery, emotions, and personal experiences (Kintsch & Rawson, 2005).

Thinking of the varying processes involved in reading, it is obvious that one definition or statement cannot capture the complexity of reading (Grabe, 2009). As Alderson (2000) suggests, reading process includes many language skills; however, some aspects of reading ability may become irrelevant or may be operationalized differently depending on the particular testing purpose. For example, the construct definitions of reading in IELTS (International English Language Testing System) and FCE (First Certificate in English) are fairly different, but they may be assessing equally valid constructs based on their purpose (Alderson, 2000).

Reading comprehension is a multidimensional and complex process. A number of factors influencing reading comprehension process have been identified in the literature. Linguistic knowledge, vocabulary knowledge, background knowledge and social factors can be counted among them.

Second language proficiency is certainly one of the factors that influence reading process. A number of studies have shown positive correlations between high L2 proficiency and better L2 reading comprehension abilities (Bossers, 1991; Lee & Schallert, 1997; Jiang, 2011). For example, Bossers (1991), in a study with 50 Turkish native speakers who learn Dutch as a second language, investigated the influences of L2 proficiency and L1 reading ability on L2 reading comprehension. The results indicated that the two independent variables together accounted for about 73% of the variance on the dependent variable, but the influence of L2 proficiency was much stronger than L1 reading ability. Lee and Schallert (1997) conducted a study with 809 Korean learners of English at different proficiency levels. The results revealed that 56% of the variance in L2 reading could be accounted for by L2 proficiency. Similarly, Jiang (2011) also investigated the role of L2 proficiency on L2 reading comprehension as a part of a study. The data from 246 undergraduate students with L1 Chinese L2 English showed that L2 proficiency is moderately correlated with L2 reading comprehension, and accounted for about 27-35% of the variance.

Background knowledge is a comprehensive factor that is closely related to reading ability. Topic familiarity and cultural background knowledge in a given subject facilitate and improve comprehension process (Carrel, 1987; Leeser, 2007; Sabatin, 2013). Leeser (2007) investigated the influence of topic familiarity on reading comprehension and found significant influence of topic familiarity on how much participants recall from the texts. Sabatin (2013), with a sample of 120 university level students, investigated the influence of cultural knowledge on reading comprehension. The results indicated that the performance of the participants who received lectures on American culture and who did not was significantly different on

reading comprehension tests administered after the lectures. Therefore, it was indicated that cultural background knowledge plays an important positive role on reading comprehension performance of students.

Formal schemata can also be counted as a part of background knowledge, and it refers to the knowledge of language, i.e. rhetorical organizational structures of different types of texts. Such knowledge is supposed to have a supportive effect on reading comprehension, and research has shown that it facilitates reading comprehension (Carrel, 1987; Zhang, 2008). In Carrel's (1987) study, for example, it was found that both familiar content and familiar rhetorical form were facilitating factors in ESL reading comprehension. Zhang (2008) compared the performance of students on three different texts with the same content but with different formal schemata - description, problem solution, compare and contrast - and found that texts with highly structured schema such as problem solution were better recalled than ones with a loose schema such as description. Familiar genres help a reader make quicker connections across bits of information in a text. For example, Rozimela (2014) examined 280 university level students in an English language study program in terms of their knowledge regarding different text genres and their performance on reading texts with these genres. She concluded that students who have higher genre awareness tended to perform better on reading tasks than the ones who have lower levels of genre awareness. Zarei and Neya (2014) found that a group of 30 Iranian EFL learners performed best on reading tasks after a discourse based instruction (register, genre and cohesive devices) when compared to other groups who received vocabulary based or syntax based instructions.

The level of vocabulary knowledge has been shown to be a determinant of the level of reading comprehension (Hsueh-chao & Nation, 2000; Qian, 2002; Zhang,

2012). Hsueh-chao and Nation (2000) investigated the influence of different levels of unknown word density on reading comprehension. They found that on average reading comprehension scores of learners predictably increase as the coverage of familiar words increases. Qian (2002) investigated the influence of vocabulary size and depth of vocabulary on reading comprehension. The results indicated that measures of both vocabulary depth and size of vocabulary significantly correlated with scores on TOEFL reading for basic comprehension. More recently, Zhang (2012), with 190 students learning English as a foreign language in a university in China, examined the relative contributions of vocabulary and grammatical knowledge on reading comprehension, and found that vocabulary knowledge was a stronger predictor of reading comprehension than grammatical knowledge.

Metacognitive knowledge, in its general sense, is the control over one's cognitive process and it is another factor affecting the reading process (Grabe, 2009). Grabe notes that metacognition involves awareness and control of planning, monitoring, repairing, revising, summarizing, and evaluating. This enables one to employ appropriate reading strategies to support comprehension. At metacognitive level, readers may consciously carry out metacognitively aware processes such as setting reading goals, making inferences in line with reading goals, monitoring comprehension, and summarizing main ideas (Grabe, 2009). McNeil (2011), based on a study with university level EFL learners with different L1 backgrounds, found that comprehension strategies, operationalized as self-questioning, better predicted reading comprehension performance than background knowledge. The study indicated that instruction on how to employ self-questioning strategies in the process of reading has a potential to improve the explanatory power of reading comprehension strategies on L2 reading. Cromley and Azevedo (2007) found that

background knowledge and vocabulary were the strongest predictors of successful reading, but use of reading strategies also had a small but significant direct contribution to reading comprehension performance.

Social and cultural factors also influence readers, both their L1 and L2 (Grabe, 2009). Grabe (2009) explains that expectations of social institutions, religion, economic status and popular culture are among these factors. For example, PISA (Programme for International Student Assessment), a large-scale assessment program, have reported that both immigrant students and students with families of low socioeconomic status have less academic success than their peers (OECD, 2009, 2012). With a meta-analysis, Sirin (2005) found that socio-economic status and academic achievement are moderately related, and especially parent's place in the socio-economic structure has a profound impact on students' academic achievement. Grabe (2009) suggests that sociocultural factors influencing L2 readers are multiplied when the dual-language mind of a L2 reader is taken into account. He states that "The social factors affecting ESL students in more advanced (post-secondary) academic settings and EFL students in language education are going to be quite different from those of L1 students from childhood to the end of high school." (Grabe, 2009, p. 169).

Taking into account different reading situations and purposes of reading, many other factors can influence reading performance such as motivation (Wang and Guthrie, 2004) and L1 reading abilities (Bossers, 1991; Lee & Schallert, 1997). However, the studies regarding reading and its relationships with other variables have mainly focused on the factors discussed above. After discussing the factors involved in the reading process in this part, the next part gives an overview of the

reading theories that attempt to explain how reading is realized and what processes are involved in it.

2.6 Theories of reading

Urquhart and Weir (1998) state that reading models can be categorized as process models and componential models. They explain that process models focus on the process of reading as an attempt to explain how various factors operate while reading takes place, while componential models attempt to explain what factors are present in the reading process, but not necessarily the interaction between them. Bottom-up, top-down and interactive models of reading can be counted as process models, and they either explain reading as sequential stages or as a non-sequential process in which various sources of information are in work simultaneously (Urquhart & Weir, 1998).

2.6.1 Bottom-up models

Influenced by behaviorism in the mid-20th century, the bottom-up model of reading posits that reading comprehension is a process in which a reader starts from decoding the smallest linguistic components and proceeds to build higher levels of meaning. Gough (1972 in Urquhart & Weir, 1998) suggests that the reader would move from decoding letters to phonemes, phonemes to words, words to sentences, and assign a meaning to the sentence. The stages are thought to be sequential and unidirectional. Grabe and Stoller (2002) explain that according to bottom-up models, comprehension comes from the information in text, and it is hardly related to the reader's background knowledge. Grabe (2009) notes that such an extreme view of reading cannot be accurate.

Bottom-up view of reading was criticized on various grounds. For example, Urquhart and Weir (1998) explained that, according to the model, processing higher-

level units should take more time than the lower level units, but this is not the case. It is possible to recognize a word more quickly than individual letters. As for the direction of processing, it has been shown that readers may use syntactic information to find the meaning of a word, which conflicts with the direction of the process urged by the bottom-up model (Urquhart & Weir, 1998).

2.6.2 Top-down models

Grabe and Stoller (2002) indicate that in top-down view, reader expectations and goals are crucial, and readers confirm or reject their expectations as they sample information from the text. Goodman (1967, in Urquhart & Weir, 1998) perceives reading as a hypothesis verification process (a psycholinguistic guessing game) in which readers start with some guesses and use the data from the text to confirm or modify their hypotheses. Therefore, top-down models are reader-driven, while bottom-up models are text-driven, and in top-down view, the reading process is assumed to be cyclical, i.e. the reader has a hypothesis, then reads the text and then turns back to their hypothesis again (Urquhart & Weir, 1998).

Schema theory also suggests that previously formed and organized knowledge guide us as we make sense of new experiences (Nunan, 1991). Therefore, schemata are important for learners in terms of utilizing linguistic cues and background knowledge in discourse comprehension. Grabe (2009) explains the schema theory as follows:

> When a word or passage activates a concept, this activation also triggers schemas - related sets of knowledge linked together in an established frame - to assist in interpreting the concept or situation and to generate inferences in support of comprehension. (p. 77)

As such, when schemas are activated, they have a supporting role on comprehension. Smith (2004) defines schemas as the representations of more general patterns or regularities that we experience. For example, when a reader reads the word

"classroom", his/her schema of a classroom enables him/her to make sense of a classroom he/she has never been before (p. 21). Smith also notes that recognizing scenes depends on the extent to which they conform to one's expectation, i.e. to the schemes he/she already has.

Rayner, Pollatsek, Ashby, and Clifton (2012) criticized top-down view of reading because they state that a great deal of evidence suggest visual processing of text is very fast; therefore, hypothesis testing and guessing behaviors cannot play a large role in reading process.

2.6.3 Interactive models

The interactive models, as Grabe (2009) explains, are based on the assumption that useful elements of bottom-up and top-down views can be combined in an interactive set of processes. Rumelhart (1977 in Urquhart & Weir, 1998) suggests that the input can be received from multiple sources, for example, orthographic, lexical, syntactic, and semantic knowledge can be all in work at the same time in reading process. Therefore, the interactive view does not accept sequential processing in reading.

Stanovich (1980, p. 36) suggests that strength in an area of knowledge or skill can compensate for a deficient area of knowledge or skill. The view draws attention to the possibility of higher level processes compensating for deficiencies in lower level processes. This approach to reading process has been known as interactive-compensatory model. Stanovich (1980) emphasizes that an interactive-compensatory model assumes the linear processing from lower level to higher level in bottom-up models is not valid (p. 36). The compensatory assumption explains that a weak knowledge or skill in one area results in greater dependence on other areas of knowledge or skill regardless of their level in the processing hierarchy.

Bernhardt (1991) also offers an interactive model. She explains that three types of variables interact in any reading activity. These variables involve linguistic, literacy and background knowledge variables, which embrace both higher level and lower level processing. Bernhardt (2011) later suggested a revised compensatory view of L2 reading, which maintains that readers use all resources from both their L1 and L2 to compensate any deficiency in the process of reading.

2.6.4 Componential models

Componential models see the reading ability as composed of discrete subskills interacting with each other (Perfetti, Landi, & Oakhill, 2005; Koda, 2007). Koda (2005) explains why adopting a componential approach is needed to understand the reading comprehension process. The first reason is the complexity of the reading process and multiplicity of the components interacting in the process. Therefore, understanding the "multilayered relationship among component skills" can enable "the identification of the sources of reading impediments" (Koda, 2005, p.19). A second reason is that componential approach can help researchers better understand the role of L2 and L1 knowledge in L2 reading and thus better understand which component skills are transferable. Furthermore, examining the components separately gives the opportunity to determine the required skills for reading proficiency because it is unlikely that all component skills are "uniformly responsible for reading ability differences" (Koda, 2005).

One of the views that assume reading is composed of component skills is Simple View of Reading. Hoover and Tunmer (1993, in Sabatini, Bruce, & Steinberg, 2013) state, "The simple view makes two claims: first, that reading consists of word recognition and linguistic comprehension; and second, that each of these components is necessary for reading, neither being sufficient in itself." (p. 3).

Grabe and Stoller (2002) explain that the basic assumption behind this view is that decoding and (listening) comprehension are together a good measure of reading comprehension. In a study, Joshi and Aaron (2000) found that the two components in this view, decoding and linguistic comprehension, account for 48% of variance in reading comprehension, but when speed of processing is also added to the model, prediction of reading comprehension improved by another 10%. However, Urquhart and Weir (1998) criticize the model in that it does not satisfactorily define these two components. For instance, they use the term "word recognition" for the process of accessing lexicon, but this could also be an important part of linguistic comprehension. Moreover, "linguistic comprehension" is operationally defined as the ability to answer questions about an oral narrative in this model, but such ability may require more than linguistic competence.

Bernhardt's (2011) compensatory model also includes three components related to reading: language knowledge, first language literacy, and other; therefore, in this sense, the model can also be seen as a componential model. Bernhardt (2011) explains that language knowledge refers to readers' grammatical competence, first language literacy refers to the overall ability to use L1 literacy in different contexts, and "other" refers to any other factors that are usually background knowledge or motivation related. Bernhardt (2011) predicts that grammatical competence can account for approximately 30% of second language reading process, and L1 literacy seems to account for another 20%. However, the rest of the variance is seen as unexplained and attributed to the "other" component.

Grabe and Stoller (2002) also offer a set of processes that they believe to be parts of reading process. They hold the view that reading comprehension is a highly complex process, and therefore, it can be better understood if analyzed in terms of

underlying processes. They explain that these processes can be categorized under two headings as lower-level and higher-level processes. Lower level processes refer to “lexical access”, “syntactic parsing”, “semantic proposition formation”, and “working memory activation”. Higher-level processes, on the other hand, refer to “text model of comprehension”, “situation model of reader interpretation”, “background knowledge”, and “executive control processes”. The researchers explain that purpose of reading will usually determine which of these processes will be emphasized. For example, in order to find simple information, word recognition abilities and some background knowledge to anticipate what to look for will be emphasized. On the other hand, reading for general comprehension will require text model comprehension and situation model interpretation. Grabe (2009) suggests that lower level processes are carried out as a part of working memory, and many aspects of higher-level component abilities are often carried out automatically except when difficulties arise.

The divisibility of reading for testing purposes is an issue of interest since 1960s (Khalifa & Weir, 2009). Regarding this issue, Khalifa and Weir (2009) state:

> If reading is divisible, examination boards would need to consider the various *components* of this ability, and account for the potentially differing non-observable mental competencies of, for example, accessing knowledge of lexis and structure, textual inferencing, building a mental model, drawing on L1 resources or integrating information within or across texts. … If reading were not a divisible construct, then examination boards might be encouraged to test only those components of reading that best met the criteria of practicality and scoring validity, for example knowledge of lexis and structure. (p. 35)

Recent research has focused more on investigating the extent to which various component skills can account for reading comprehension. For example, Oakhill, Cain and Bryant (2003) investigated the relationship of a set of skills to reading comprehension. They found that the subskills that most significantly accounted

for reading comprehension were comprehension monitoring, text integration skill, and story structure knowledge. Farhady and Hessamy (2005) used exploratory and confirmatory factor analysis on data from 1606 EFL learners in order to investigate variables underlying reading ability. They developed a test of reading in an attempt to operationalize 28 subskills of reading which they specified based on previous research. The results showed that L2 reading ability is composed of a number of underlying macro-skills such as inferential and interpretive skills, linguistic and textual contributory skills, understanding explicit information, and process analysis. Nassaji (2003) investigated how higher level syntactic and semantic processes and lower level word recognition, phonological and orthographic processes contribute to reading comprehension. Each component was measured on different tests and it was found that they all had significant positive correlations with reading comprehension, and they all contributed significantly to the discrimination between high achieving and low achieving ESL readers.

On the other hand, Weir and Porter (1994) cite Rost (1993) who found evidence regarding the "unidimensionality" of reading through factor analysis from a sample of native speakers. Rosenshine (1980 in Khalifa & Weir, 2009) examined the previous studies to look for empirical evidence for divisibility of reading comprehension skills. The review of these studies indicated that different studies found different underlying subskills. Since the results across studies were inconsistent, it was concluded that there is no clear evidence for the divisibility of reading comprehension.

2.7 A reading model by Khalifa and Weir

According to Khalifa and Weir (2009) reading starts with a purpose for reading. Based on this purpose, reading takes place in different types and at different levels, and in the course of reading, readers make use of various sources of knowledge. Urquhart and Weir (1998) define types of reading that one employs based on their purpose in a similar fashion to the model by Khalifa and Weir (2009). Therefore, before moving to the reading model under discussion, Urquhart and Weir's (1998) explanations regarding types of reading will be briefly mentioned.

As stated previously, Urquhart and Weir (1998) briefly define reading as dealing with language messages in written or printed form. They argue that reading includes different strategies with different dimensions: local and global level text reading; expeditious and careful reading strategies. Local level comprehension refers to understanding of "micro-level structures" of the text such as the meaning or function of words, phrases or sentences. Global level comprehension involves comprehending the "macro-structure" of the text, which can include main ideas or discourse topic. Careful and expeditious readings represent different strategies that a reader adopts according to different reading purposes. Careful reading refers to paying attention to details and attempting to handle the majority of information in the text. On the other hand, expeditious reading involves quick and efficient examination of the text. Expeditious reading can involve reading for gist, locating specific information or reading selectively to achieve a specific goal.

According to Urquhart and Weir's (1998) reading model, careful reading at the global level requires comprehension of the majority of text such as reading for study. On the other hand, careful reading at the local level involves focusing on the local parts of a text such as predicting the meaning of a word based on its content or

understanding lexical or grammatical cohesion. Search reading and skimming are expeditious reading at the global level. While skimming refers to reading for main ideas and discourse topic, search reading refers to quickly locating relevant information. Lastly, expeditious reading at the local level, which is associated with scanning, refers to reading to locate a specific word, figure, etc. Readers may adopt these different reading types based on their purpose (Urquhart & Weir, 1998).

Khalifa and Weir's (2009) model also explain reading as taking place in various types and at different levels. However, this revised model is based more on processing, and accounts for the interactions between reader's purpose, core cognitive processes, and knowledge stored in long term memory (Ünaldı, 2010). Figure 1 illustrates the reading model offered by Khalifa and Weir (2009). The goal setter on the left side is associated with deciding on the purpose of reading. Purpose of reading is important since decisions based on the purpose determine which processes will be more important in the central core of the model (Khalifa & Weir, 2009). The column in the center indicates the cognitive processes and it is hypothesized that difficulty in reading is a result of the level of the processing required (Ünaldı, 2010). The column on the right shows the sources needed at different levels of processing.

Khalifa and Weir explain that monitor, on the left column, refers to self-monitoring oneself in the process of reading, and it is activated based on reader goals. By self-monitoring, readers may decide to change the type of reading they adopted, check word recognition or syntactic parsing, or determine how successful their understanding of argument structure of the text is.

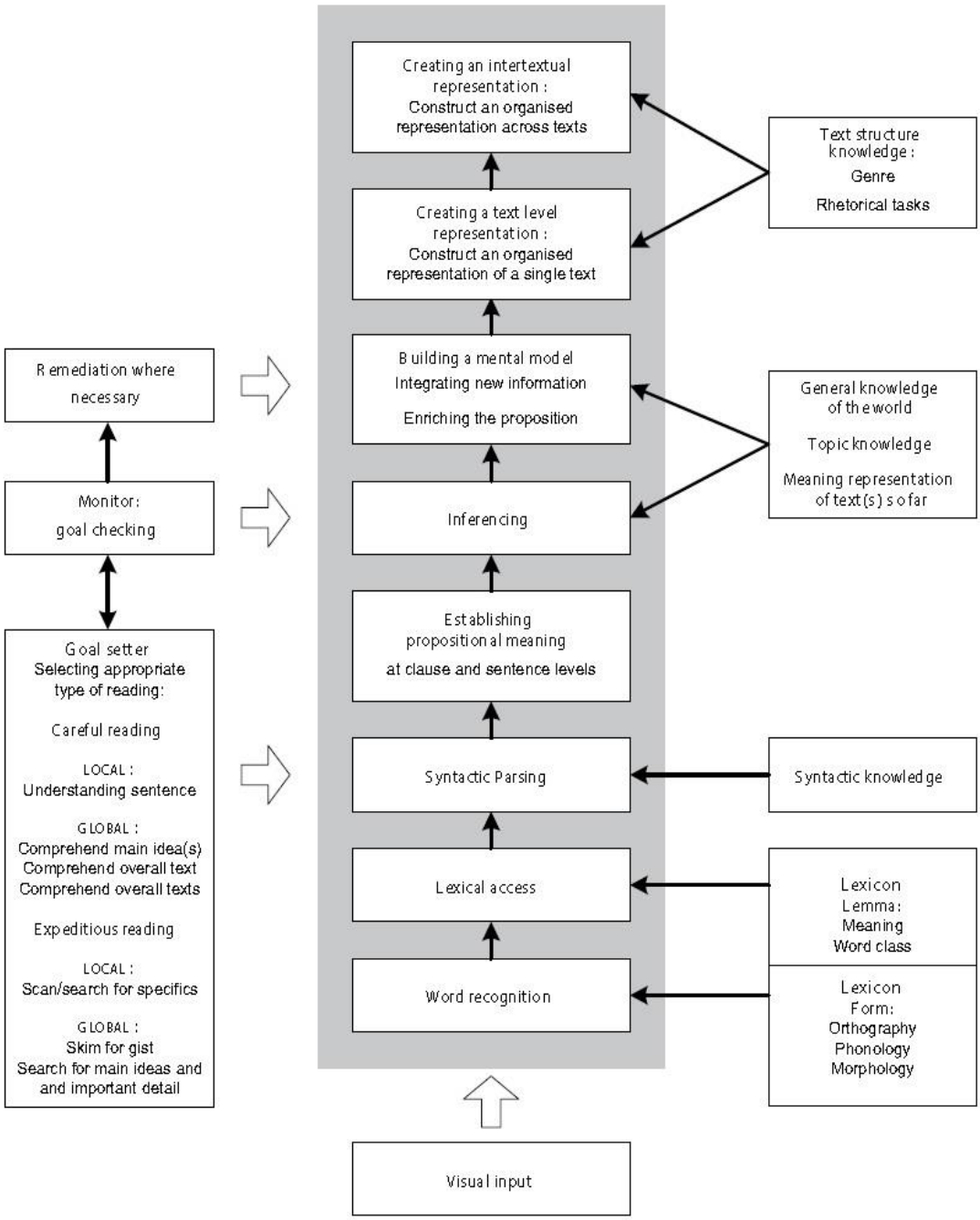


Figure 1. The reading model illustrated by Khalifa and Weir (2009, p. 43).

The types and levels of reading are also pictured on the left column of Figure 1. According to the model, in careful reading where the aim is to comprehend the complete meanings suggested in the text, readers may work at global or local level. At global level, readers try to build up an understanding of the text as a whole based on the majority of the information presented in the text. In careful global reading, the whole text is read relatively carefully, and readers may need all the processes in the

central core of the model, finally creating a discourse level structure. Khalifa and Weir explain that careful reading at the local level is associated with processing at the decoding level to establish a basic understanding of a proposition. It may require inferencing at sentence level, but it does not require integrating information across local parts of the text. Since careful local reading entails establishing propositional meaning at sentence level, it also requires the processes in the central core below this level, i.e. word recognition, lexical access and syntactic parsing.

Expeditious reading refers to reading quickly, selectively and efficiently to reach the desired information (Weir & Khalifa, 2008; Khalifa & Weir, 2009; Urquhart & Weir, 1998). Expeditious reading includes skimming, scanning and search reading, and like careful reading can be carried out at local and global levels. As also defined in Urquhart and Weir's (1998) model, skimming is reading to obtain gist or general impression of a text, so readers try to build a macro-structure of the text. Khalifa and Weir (2009) explain that scanning is reading selectively at word level to find specific elements such as a figure or name; therefore scanning mainly requires the accurate recognition of word or words. Search reading is carried at local level when the target information to be located is within a sentence, and at global level when the information needs to be put together across sentences. They also explain that in search reading readers can make use of central core processes up to building a mental model, but creating text level structure will not be needed. Urquhart and Weir (1998) and Khalifa and Weir (2009) note that although we frequently employ scanning, skimming and search reading in the real world, usually careful reading has been the focus of teaching and testing reading, which means expeditious reading has been ignored. This claim has been supported by a number of recent studies (Devi, 2011; Katalayi & Sivasubramaniam, 2013).

Weir and Khalifa (2008) hypothesize the order of difficulty between types of reading in their model. They present the following list which starts with the easiest and ends with the most difficult one:

1. Scanning/search reading for local information
2. Careful local reading
3. Skimming for gist
4. Careful global reading for comprehending main idea(s)
5. Search reading for global information
6. Careful global reading to comprehend a text
7. Careful global reading to comprehend texts (p. 9)

The researchers also add that it should be possible to claim 2 is more difficult than 1 and 7 is more difficult than 6, but the ones in the middle are closer to each other in terms of difficulty. Therefore, contextual parameters might come into play to establish difficulty differences between the three skills in the middle (Weir & Khalifa, 2008).

Khalifa and Weir (2009) draw attention to a number of contextual parameters in terms the cognitive load that might influence performance in reading. Context validity is both related to task setting and linguistic demands. Task setting includes issues such as response format, weighting, knowledge of criteria, order of items, channel of presentation, text length, and time constraints, while linguistic demands include issues such as discourse mode, reader-writer relationship, functional resources, grammatical and lexical resources, and content knowledge. In accordance with Khalifa and Weir's (2009) explanations, these parameters are discussed below.

The response format of a task can either require selecting the answer from a set of options or producing the answer. Multiple choice, true false and matching items are examples of selected response format. Khalifa and Weir (2009) explain that multiple choice format seems to be less representative of real life tasks, but it is closer to activating the natural processing for careful and expeditious reading (p. 84). They also explain that well-constructed multiple choice items tend to be efficient

discriminators between achieving and non-achieving test takers. Therefore, such items should be prepared with great care. For example, options should be constructed carefully in order to avoid giving unintended clues for the correct response (Haladyna, Downing and Rodriguez, 2002) and answers should not be open to subjective judgment (Chen, 2010). Examples to constructed response format include short answer, information transfer, and cloze tests. Khalifa and Weir (2009) suggest that although short answer format can elicit skimming, scanning or search reading, the involvement of writing may pose problems; therefore, the range of possible answers and the amount of writing should be limited. For example, in order to avoid the interference caused by writing, information transfer can be employed where candidates label a diagram or complete a chart based on a text (Khalifa & Weir, 2009). It is also stated that cloze test can reflect only a limited part of reading proficiency because it runs the risk of focusing on only local parts of the text, and may not require global understanding. Afflerbach (2011) indicates that generally constructed response items are regarded as more demanding when compared to multiple choice items.

Weighting is about assigning different amounts of score to different items. For example, extracting main ideas is probably more important than finding specific details. Khalifa and Weir (2009) emphasize that weighting should be based on a rationale and candidates should be informed about it.

Knowledge of criteria refers to the clear communication of judgement criteria to candidates. Khalifa and Weir (2009) note that candidates should be informed, for example, whether they will be scored for the accuracy of their responses regarding a set of comprehension questions. This parameter is more related to tests that involve constructed item format.

Order of the items is suggested to follow the order of processing the text, especially those requiring careful reading (Khalifa & Weir, 2009). On the other hand, in expeditious reading the order can be preferably mixed because readers are expected to work at local level or not expected to have a thorough understanding of the text. Although it is not possible to predict how a reader will approach a given text, it seems a good idea to present items aiming careful global reading after the ones aiming local level reading or expeditious reading (Khalifa & Weir, 2009).

Channel of presentation, as explained by Khalifa and Weir, is about the existence of information that is not in the written form such as pictures or tables in a test. Information presented in more than one channel is supposed help readers encode information more easily. For example, seeing pieces of relevant information in a diagram can be easier than integrating them across paragraphs.

Text length and time constraints are two other measures to take into account regarding task setting. Afflerbach (2011) states that text length and complexity usually increase with the level of the test. It is stated in CEFR that "in general a short text is less demanding than a long text on a similar topic as a longer text requires more processing and there is an additional memory load" (p. 166). Similarly, Khalifa and Weir (2009) suggest that longer texts and sentences are more challenging in terms of both lower and higher level processing, and it is possible to elicit more types of reading with longer texts. Decisions regarding time constraint influence the type of processing and hence the reading strategy readers will adopt (Khalifa & Weir, 2009). Within this scope, the speed of reading is supposed to be different in expeditious and careful reading strategies (Hughes, 2003; Khalifa & Weir, 2009); therefore, especially at higher levels of proficiency, time should be appropriately

constraint to elicit expeditious reading. Otherwise, readers may adopt careful reading in the abundance of time where careful reading is not intended.

Discourse mode is regarded as a parameter influencing the linguistic demands of a text. Discourse modes or rhetorical organisations such as problem/solution and cause/effect have been found to be better comprehended by readers when compared to descriptive ones (Khalifa & Weir, 2009). For example, Meyer (1975, cited in Alderson, 2000) found that the same paragraph was recalled better when presented as a solution than when appeared as an item in a list. Therefore, she concluded that readers may find the organisation of some texts easier to follow than others. More recently, Jaliehvand and Moses (2014) compared two groups of EFL students' performance on two texts with the same content but with different rhetorical organisations (descriptive and causative). They found that students who read the text in causative organisation performed significantly better than the ones who read the text in descriptive organisation. Zhou (2011), in a study with 133 Chinese advanced EFL learners, found that students performed significantly better on expository texts when compared to narrative texts.

Reader-writer relationship issue signifies that the anticipated reader group of a text will affect the discourse of the material (Khalifa & Weir, 2009). The amount of information and specificity level of content in a text will have different impacts on reader groups who share the same linguistic and content knowledge of that discourse and who do not (Khalifa & Weir, 2009).

Functional resources refer to functions of the text such as recommending, justifying, informing or disagreeing. Khalifa and Weir (2009) explain that some basic functions such as understanding opinions can be expected at any level of proficiency while other functions such as hypothesizing will not be expected until higher levels

of examination. The CEFR also basically provides what language functions can be carried out at different proficiency levels by describing what learners can do at each level. For example, the statement "Can identify the main conclusions in clearly signalled argumentative texts" clearly means that the readers at this level can read simple texts that have the functions of informing and justifying.

Grammatical and lexical resources are about syntactic and lexical complexity, which have an impact on linguistic demands of reading texts. Texts with less complex grammar and more frequent words tend to be less demanding (Khalifa & Weir, 2009). In a study with 64 Hungarian native speakers learning English, Morvay (2012) found that L2 lexical knowledge and L2 reading comprehension had a significant positive correlation. The results also indicated that L2 syntactic knowledge was a statistically significant predictor of L2 reading comprehension. Khalifa and Weir (2009) provide of an overview of gradually more complex grammatical structures that are expected on five Cambridge ESOL examinations. These examinations are thought to be aligned with the upper five proficiency levels on CEFR (i.e. excluding A1). For example, the examination for A2 level is supposed to involve normally simple sentences while B1 level examination is supposed to involve mainly simple sentences but occasional use of relative and other subordinate clauses. Regarding lexical complexity, the frequency of words seems to be a good criterion to arrange complexity level (Khalifa & Weir, 2009). However, word frequency is still an indirect indicator of text complexity. Khalifa and Weir (2009) explain that Cambridge ESOL examinations generally involve reading texts that include more vocabulary out of the first 1000 and 2000 word lists as the aimed proficiency level of the examinations increases. Hsueh-chao and Nation (2000) devised a number of texts each of which included different amounts of unknown

words (non-words) in order to investigate what percentage coverage of text is needed for unassisted reading. Based on the results, they concluded that comprehension scores increase as the coverage of known words increase, and readers need to know around 98% of the words in a text for uninterrupted reading.

Finally, regarding content knowledge, Khalifa and Weir (2009) warn that "propositional inferencing" is considered acceptable, but "pragmatic inferencing" is not, which means inferencing should be by means of the information in the text, not reader's prior knowledge. Alderson (2000) states that absence of background knowledge about a given text should not put a group of test takers in a disadvantaged position (p. 29). Urquhart and Weir (1998), on the other hand, advise that testers should avoid texts that are so unfamiliar to candidates. Khalifa and Weir (2009) note that in order not to upset certain groups of test takers, Cambridge ESOL examinations tend to choose neutral topics such as health, travel, weather, sports, arts, and education, and avoid topics such as war, politics, religion, historical references, terminal diseases, natural disasters, and common phobias (p. 139).

2.8 Conclusion

Validity and reliability are two major issues related to testing. There are a number of approaches that have been used by researchers and test developers to look for various sources of evidence for validity and reliability. Such evidence is certainly needed to assess whether score on a test is consistent and fits to its purpose. Only in that way it is possible to decide how much confidence to put in the decisions made based on the scores.

Developing a test is an iterative process. It starts with a perceived need for the test, and continues with planning the development process and operationalizing the constructs through specifications. The feedback from experts in the development

process and the data extracted from trials are invaluable to improve the test. The new data from administrations of the test may always urge improvements and modifications.

Reading is a complex construct and theories explaining reading have changed over time. The behaviorist explanations about reading were opposed by top-down views which take into account the reader factor. However, recently more interactive explanations to reading that draw on both bottom-up and top-down views have usually been embraced. There has also been research on the divisibility of reading ability. Such studies have usually investigated sub-constructs or subskills that can be underlying the general reading comprehension skill.

The reading model proposed by Khalifa and Weir (2009) suggest that reading is carried out either expeditiously or carefully, and this takes place at either global or local levels. The type of reading employed and cognitive processes needed are closely dependent on the purpose of reading. The researchers also draw attention to a number of contextual features that might influence reading performance.

After this chapter that reviewed the relevant literature, the next chapter describes the methods used to investigate the research questions.

CHAPTER 3

METHODOLOGY

3.1 Introduction

The aim of this study is to investigate the item characteristics, validity and reliability features of the five reading tasks developed for learners of Turkish as a foreign language. To address the aim of the study, both quantitative and qualitative techniques were employed. Detailed information related to participants, instruments and the techniques can be found below.

3.2 Participants

The participants of the study were 62 students who were studying at Boğaziçi University, Turkey at the time through Erasmus, which is an international student exchange program between universities in Europe. Their age ranged from 19 to 31, with a mean of 23. Most of them had arrived in Turkey one and a half months earlier than the time data were collected. They came from many different countries (see Table 2). Based on the participants' reports, their average duration of stay in Turkey was 14 months (*SD*=31.5) and their average length of learning Turkish was 59 months (*SD*=86.61). The students were at Boğaziçi University to spend one or two semesters and then to return to their home universities. All of the participants were taking Turkish for Foreigners (TKF) classes, 211, 315 and 317 offered by the Department of Turkish Language and Literature. The course instructors reported that the proficiency level of the students taking TKF 211 could be considered intermediate while those taking TKF 315 and 317 could be considered advanced. The instructors indicated that the placement of students to different classes is carried out by the Turkish Language and Literature Department based on the interviews conducted with each student. However, the students' preferences regarding which

course to take, which is usually influenced by their perceived level of proficiency, is also considered.

Table 2. Number of Participants Coming from Each Country.

Nationality	Frequency	Percentage
Germany	17	27.4
Japan	9	14.5
USA	7	11.3
Greece	4	6.5
Cyprus	3	4.8
Netherlands	3	4.8
France	2	3.2
Jordan	2	3.2
Azerbaijan	2	3.2
Italy	1	1.6
Serbia	1	1.6
Austria	1	1.6
Great Britain	1	1.6
Kosovo	1	1.6
Mauritius	1	1.6
Norway	1	1.6
Sweden	1	1.6
Turkey	1	1.6
Turkmenistan	1	1.6
Iran	1	1.6
Iraq	1	1.6
Syria	1	1.6

A learner profile form (see Appendix A) was developed to ask the participants to evaluate their linguistic ability in Turkish according to the Common European Framework of Reference (CEFR) scale. The participants' self-evaluations indicated that TKF 211 group's average proficiency level was 2.69 where 2 means A2, and 3 means B1 on the CEFR scale. On the other hand, TKF 315/317 group's average proficiency level was 4, which corresponds to B2 level on the CEFR scale. Unfortunately, objective evidence related to the proficiency level of the students in Turkish is nonexistent. Those who were taking TKF 211, 315 and 317 courses were intentionally chosen to be the participants of the study because the students in lower-

level TKF classes, such as 111 or 112, were predicted not to have sufficient Turkish proficiency to complete the tasks employed for study.

3.3 Instruments

The following instruments were employed in this study.

3.3.1 Reading tasks in Turkish

The reading tasks were developed by the researcher through a process of text selection, item writing, consulting experts and item trial. Information regarding the fields and topics of the texts used in the tasks are presented in Table 3 below. In total, 6 reading tasks, with different intended levels from B1 to C2 on the CEFR scale, were developed and 5 of them were administered. Each reading task was administered to 31 participants. The tasks are presented in Appendix B.

Table 3. Text Topics.

Text	Field	Topic
1	Environment	Köpekbalıklarının Soyu Tükeniyor (Sharks are Going Extinct)
2	Cinema	Hoffman'a Veda (Goodbye to Hoffman)
3	History	Lale (Tulip)
4	Literature	Beyoğlunun En Güzel Abisi (When Pera Trees Whisper)
5	Biology	Yaprağın Yapısı (Structure of Leaf)

While developing the tasks, based on the reading theory by Khalifa and Weir (2009), a list of reading skills that were thought to be relevant to careful and expeditious reading strategies at global and local levels was prepared by an expert. In addition, informed by the Can-do statements in the CEFR for each proficiency level, item specifications were developed by sampling the appropriate reading skills that are thought to be relevant to each level. In the design of the item specifications, the information regarding contextual features of the tasks was presented as guided by

Ünaldı's (2010) study. The item specifications regarding each task were later revised based on study results and are presented in Appendix C.

3.3.2 CEFR

CEFR is a reference for curriculum development, teacher training, and assessment. It includes a number of scales describing six levels of proficiency. Can-do statements of reading do not provide a theory of development of reading abilities; however, they provide a taxonomy of behaviors (Alderson et al., 2004). Therefore, in this study, it was aimed to merge the reading model by Khalifa and Weir (2009) with the taxonomy of CEFR. The skills operationalized in the test tasks were chosen by comparing this reading model and CEFR Can-do statements. The researcher and an expert decided on the level of performance that can be expected from test takers in relation to these skills. It was decided that certain skills can be carried out only by test takers at or above certain levels. For example, the skill "Distinguishing fact from opinion" was assumed to require careful local reading and was thought be relevant from B2 level onwards depending on the Can-do statements on B2 level: "Can obtain information, ideas and opinions from highly specialized sources within his/her field, and can read articles and reports concerned with contemporary problems in which the writers adopt particular attitudes or viewpoints." On the other hand, the skill "Retrieving specific information by scanning text" was assumed to require expeditious local reading by definition and was thought to be suitable for all levels between B1 and C2 depending on the Can-do statement on B1 level: "Can scan longer texts in order to locate desired information."

3.3.3 Other tests

The listening, speaking and writing tests, from which the scores were correlated with the reading tasks, were being developed by other researchers at the time of this study

and administered to the same group of test takers. The listening test included five tasks aimed for different proficiency levels from A1to C1. Based on audio recordings, the test takers were expected to carry out tasks such as answering a set of multiple choice items or completing blanks in sentences. The speaking test included six tasks, two of which required test takers to interact with each other. Finally, the writing test had two tasks aiming B1 and B2 level test takers. One of the tasks required information transfer from graph while the other one was argumentative and was initiated with a question.

3.3.4 Reading test expert evaluation form

Following Bachman's (2004) suggestion, a rating scale (see Appendix D) was developed by an expert in order to extract evaluations of expert judges with the aim of looking for content validity evidence. The rating scale included propositions related to the abilities being tested and other information such as text topics and item characteristics, and the experts were expected to indicate their agreement on a four-point scale. One of the experts is a professor of Applied Linguistics at Boğaziçi University, and she has delivered Turkish courses for foreigners for over ten years. The other expert holds a PhD in the area of testing languages with expertise in test validation. The aim of the expert ratings was to seek for evidence for content validity by investigating the degree to which the judgments are parallel with the test specifications.

3.3.5 Learner profile form

A learner profile form (see Appendix A) was developed to collect information related to the learners' demographics such as age, nationality, mother tongue, and so forth. It also aimed to extract information about learners' duration of stay in Turkey

and duration of exposure to Turkish, as well as what level of proficiency they believe themselves to be at on the CEFR scale with descriptors.

3.4 Procedure

The data from the participants and from experts were all collected in the fall semester of 2014. The tasks were administered to the participants within a time span of two weeks to ensure that learning would not be a confounding factor. The participants were given the tasks in the classroom environment and they were expected to complete the tasks within the allotted time. Because of time limitations, the test was delivered as two forms in each class: Form A consisted of reading tasks 1, 2 and 5 while Form B consisted of tasks 3 and 4. All of the participants completed two tasks or three tasks depending on the form they took. Each task was delivered separately. The participants were informed that the tasks belonged to the reading component of a test of Turkish, which was in progress. Table 4 summarizes the number of participants completing each task and the allotted time for each task. The participants were not provided with any further explanation related to the tasks or individual items since the explanations related to tasks were presented as instructions written on the task sheets.

Table 4. Number of Participants and Duration of Reading Tasks.

Task	Intended level	Number of students who completed the task	Time given for the task
1	B1	31	15 min.
2	B2	31	15 min.
3	C1	31	20 min.
4	C2	31	20 min.
5	C2	31	10 min.

3.5 Data analysis

Based on the data from the test takers' performance on the tasks, frequencies, item analyses and correlational analyses were carried out using the software program SPSS 20.0. Distractor analysis and analysis of expert ratings were conducted using Microsoft Excel 2010.

Research question (RQ) 1: Do the experts agree on the operations measured by the test items as specified by the test writers?

RQ 1 is investigated by collecting expert opinions and analyzing their agreement both with the intended reading operations and with each other on these operations. For this purpose, data from the Expert Opinion Form were analyzed by calculating percentages of agreement. The appropriateness of instructions, items and texts were also analyzed through Expert Opinion Form. Verbal feedback was also taken on these issues.

RQ 2: Do the test tasks differentiate between higher and lower proficiency groups?

RQ2 is addressed by performing a *t*-test to find out whether the participants from different proficiency levels had significantly different sets of scores on the tasks. A primary expectation from a given test is that it should discriminate between students with higher and lower levels of knowledge. Students with higher and lower levels of knowledge are determined based on a criterion. For the same concerns, the performance of the two groups – one that takes TKF 315 and/or TKF 317 classes and one that takes TKF 211 classes – were compared to investigate whether TKF 315/317 group performed better than TKF 211 group. Generally, the students in TKF 315/317 are expected to outperform those in TKF 211 because they are supposed to be at a higher level of Turkish language proficiency. With small samples, generally

up to 30 subjects in each group, *t*-test is a frequently employed statistical analysis to explore the differences between two samples. One assumption of *t*-test has been reported to be that the two samples should be normally distributed. However, Bachman (2004, p. 236) states that although such an assumption is cited in many textbooks on statistics, violation of normality assumption has been shown to have nearly no influence on the results when using the two-tailed *t*-test. Another assumption of the *t*-test is the homogeneity of variances between the two groups, which is usually indicated by a statistical figure that comes from Levene's Test for Equality of Variances or by an F-ratio that comes from F-test. The final assumption of *t*-test is the independence of the observations, which means the performance of each group and individual test takers should not influence each other in any way. These assumptions were checked before the *t*-test analysis was carried out.

RQ 3: What are the psychometric characteristics of the items for each reading task?

a. What levels of item difficulty are reflected by the scores of the test takers?

b. To what extent do the items discriminate between test takers' reading abilities?

c. What are the internal reliability values for each task?

RQ 3 is investigated through item analyses in which difficulty and discrimination values of items are explored. Internal consistency of items and the efficiency of distractors are also analyzed for research question three. Item difficulty stands for the percentage of test takers who answered the item correctly. Item difficulty can take a value between 0 and 1, and higher values mean the item is easier. Too difficult or too easy items are unfavorable (Alderson et al., 1995). Since this test is an early version of a reading test, the limits for item rejection were set at

0.20 and 0.80 boundaries, following Bachman (2004). Item discrimination value indicates how efficiently an item discriminates between test takers with higher and lower levels of knowledge. It can be investigated by calculating point biserial correlation coefficient which is based on the correlation between single items and the total test scores (Bachman, 2004). Therefore, discrimination values are computed based on this correlation in the present study. Point biserial correlation (or item-total correlation) coefficient is advised to be over .30 by Bachman (2004). Since this is an early version of the test with a limited number of subjects, items with an item-total correlation value 0.20 and above will be accepted, following Brown (2005). Internal reliability is usually calculated by Cronbach's alpha which is based on the correlations between items. Items that assess the same or similar constructs are expected to have high intercorrelations (Alderson et al., 1995). Cronbach alpha values indicate how internal consistency of a test would change when an item is excluded from the test.

Distractor efficiency analysis investigates how efficiently the distractors of a particular item do their job. The distractor efficiency analysis in the present study is based on the percentage of responses that each option draws on each item. Bachman (2004) states that each distractor should draw at least 10% of the responses. If it does not, this means it is not working. Taking into account the limited number of subjects in the present study, each distractor is evaluated whether it drew at least one response which equals to 3.22% of all responses from 31 subjects in our case.

RQ 4: Do scores on the reading test correlate with the scores obtained from listening, writing and speaking tests administered to the same group?

RQ 4 is addressed by investigating the relationship between four sets of scores that come from the four sub-components of the test. The relationships between the sub-tests are calculated as Pearson product-moment correlation coefficient.

The next chapter presents the results from the analyses explained above.

CHAPTER 4

RESULTS

4.1 Content validity

The data from expert judgments regarding appropriateness of text topics, the abilities tested by the items, and item characteristics are presented below. Basic statistics related to text characteristics are summarized in Table 5 in order to give an idea about the readability of the texts.

4.1.1 Expert ratings for the instructions, questions and text of tasks

The experts evaluated the wording and quality of instructions, questions and texts related to each task. The rating scale had the following statements which were expected to be rated by the experts on a four point scale: 1 - strongly disagree, 2 - disagree, 3 - agree and 4 - strongly agree. The ratings provided by the experts are presented in Table 6 below. Ratings over 2 signify that the expert agrees with the related statement, while ratings 2 and 1 mean that the expert does not agree with the related statement. The rating scale originally does not have half-point options, but one of the experts preferred using half-point ratings as shown in Table 6.

For Task 1, Rater 1 did not indicate any problems related to the instructions; however, Rater 2 indicated that the instructions were not clear and adequate. A further written feedback from Rater 2 suggests that the instruction at the beginning of the task should include the information of how to read the text, i.e. carefully or quickly. Therefore, also taking into account the theory which the test is based on and which explains the different functions of careful and expeditious reading styles, such wording was employed for the elaboration of the instruction related to Task 1. A second problem indicated by Rater 1 related to Task 1 is that the wording of the questions was not easier than the text. Thus, a simplification on the wording of the

Table 5. Text Characteristics.

Text	Text type	Intended Level	Characters	Words	Sentences	Paragraphs	Characters per word	Words per sentence	Sentences per paragraph
1	expository	B1	2668	371	26	7	7,1	14.27	2.6
2	expository	B2	2020	335	22	4	6	15.23	5.5
3	expository	C1	2833	421	30	7	6,7	14.03	4.3
4	narrative	C2	2924	450	33	3	6,5	13.64	11
5	descriptive	C2	1146	180	16	5	6,4	11.25	2.7

Table 6. Experts' Ratings Regarding Instructions, Questions and Texts.

Related part of the task	Related statement on the rating scale	Rater 1					Rater 2				
		T1	T2	T3	T4	T5	T1	T2	T3	T4	T5
Instructions	Instructions are clear.	4	4	3.5	4	4	2	1	4	4	1
	Instructions are adequate.	4	4	3.5	4	4	2	3	3	4	1
	Instructions are relevant.	4	4	3.5	4	4	4	2	4	4	1
Questions	The questions are clear.	3	4	3.5	4	4	3	2	2	4	1
	The language of items is easier than the language of the text.	2	2	4	4	4	4	3	3	3	
	The questions can only be answered if the text is read.	4	4	4	4	4	3	4	3	4	2
Text	The text is appropriate in an academic context.	3	2.5	4	2.5	4	4	2	2	3	4
	The text length is appropriate in an academic context.	4	2.5	4	4	3.5	4	2	3	3	3
	The text does not require high levels of knowledge to comprehend.	3	4	3	3	3	3	1	3	4	3
	The text does not require cultural knowledge to comprehend.	2.5	2.5	2	2	4	4	1	3	2	4

Note. T = Task.

questions was suggested wherever possible on Task 1.

Rater 1 expressed the same concern also for Task 2; therefore, a similar revision to simplify the language of items was suggested for Task 2, as well. Rater 2 suggested that the instructions and questions might be difficult to understand because the task is a sophisticated one, so they could be worded simpler if possible. Moreover, Rater 2 also suggested that the length of the text should be revised because test takers at that intended level may find it short.

Related to Task 3, Rater 2 suggested that questions might not be clear. The written feedback by Rater 2 indicated that a small alteration in the wording of the prompt sentence of Item 7 was needed. As the focus of that sentence should be on the unpredictability of the diversity of people who are interested in the flower in question, rather than the number of people. Rater 2 also suggested that the multiple-choice question format of Item 8 should be altered because the options may function

as clues, which can be a confounding factor when testing a specific reading skill. Another suggestion by Rater 2 was the alteration of the wording of a distractor of Item 1 and a distractor of Item 4. The suggestions by Rater 2 were executed on Task 3 in the revised version.

Regarding Task 4, Rater 2 suggested a minor change on the wording of a distractor of Item 3, which, she believed, would make the item clearer. As for Task 5, Rater 2 indicated problems related to the clarity of instructions and questions. This judgment was reinforced by a closer inspection of the answers provided for Task 5. It was revealed that several participants did not understand the instructions because they provided answers which were not aimed at all.

In the commentary section of Expert Judgment Form, Rater 2 provided written feedback and she expressed reservations about the appropriateness of Text 2 in an academic context, and whether it could be biased because of a group's background or cultural knowledge. She explained that the content of the text might be favoring those who are familiar with the cinema terminology. Related to Text 3, Rater 1 expressed a similar concern that the text might be culturally biased, which can be because of the vocabulary used in the text as it includes several nouns that can be related to the Ottoman culture. The ratings by both experts related to Text 4 indicate the same one concern which is the probability of cultural knowledge interfering with the comprehension of the text. Since Text 4 is intended for C2 level test takers, cultural terminology might be tolerated up to a certain level; however, the text can be revised to replace the words that might require cultural knowledge, and adjusted to a level where it entails cultural knowledge at the minimum level.

4.1.2 Expert ratings for the skills measured by items

On the rating scale, the following skills (see Table 7) which are thought to be relevant to the five reading tasks were listed. For the items of each task, the raters marked the skills that they thought the item aimed to measure.

Table 7. List of Skills.

Reference number	Skills
1	Skimming for overall gist
2	Demonstrating understanding of text as a whole
3	Identifying topic of text
4	Identifying function of text
5	Distinguishing main points of text from subsidiary ones
6	Retrieving specific information by scanning text
7	Locating and selecting relevant factual information to perform task
8	Demonstrating understanding of how text structure works
9	Distinguishing fact from opinion
10	Deducing meaning from context
11	Interpreting text for author's attitude and style
12	Making inferences from information given in the text
13	Making use of clues such as subtitles, illustrations

Tables 8 - 12 summarize the specific purpose of each item that corresponds to skills on Table 7, and the ratings by the experts. Any parallelism between the two raters and the corresponding skills on the rating scale suggests that a specific item is more likely to measure the skill that it is supposed to measure. Since a hundred percent parallelism is hard to achieve, discrepancies both between the two raters and between the intended skills and raters are expected. Such discrepancies may provide valuable feedback in order to revise the items.

Table 8 suggests that although skills 6 and 7 overlap for Item 1, skill 4, 10 and 12 were not intended at all, but Rater 1 thought 10 and 12 as relevant, and Rater 2 thought 4 as relevant. Similarly, skill 12 in Item 2 was not evaluated as relevant by raters; moreover, Rater 1 indicated skills 1 and 10 as relevant while they were not aimed. Table 8 also suggests that although Rater 1 thought it relevant, skill 10 was not intended for Item 3, Item 4, and Item 5. Rater 1 also thought skills 1 and 5 as relevant to Item 6.

Table 8. Experts' Ratings Related to Task 1.

Item	Intended purpose of the item	Corresponding skills on the rating scale	Rater 1	Rater 2
1	Identifying explicit details from the text	6, 7	6, 7, 10, 12	4, 7
2	Making inferences and drawing accurate conclusions based on explicit information from the text	6, 7, 12	1, 6, 7, 10	7
3	Identifying explicit details from the text	6, 7	6, 7, 10	7
4	Identifying explicit details from the text	6, 7	6, 7, 10	7
5	Identifying explicit details from the text	6, 7	6, 7, 10	7
6	Identifying the author's purpose based on the explicit and implicit information from the text	2, 3, 4, 11	1, 2, 3, 4, 5	2, 3, 11

Regarding the skills assessed in Task 2, Table 9 shows that the highest disagreement seems to be about Item 1. Although the raters agreed on the intended skills, they also indicated extra skills that may also be relevant to the item.

Table 9. Experts' Ratings Related to Task 2.

Item	Intended purpose of the item	Corresponding skills on the rating scale	Rater 1	Rater 2
1	Distinguishing between fact, opinion and non-existent information in the text	7, 9	2, 4, 5, 6, 7, 10	7, 9, 12
2	Distinguishing between fact, opinion and non-existent information in the text	7, 12	7	7, 12
3	Distinguishing between fact, opinion and non-existent information in the text	7	7	7, 12
4	Distinguishing between fact, opinion and non-existent information in the text	6, 7, 9, 10	6, 7, 10	7, 9, 10, 12
5	Distinguishing between fact, opinion and non-existent information in the text	6, 7, 9, 10	6, 7	7, 9, 10, 12
6	Distinguishing between fact, opinion and non-existent information in the text	6, 7, 9	6, 7	7, 9
7	Distinguishing between fact, opinion and non-existent information in the text	6, 7, 9	6, 7	7
8	Distinguishing between fact, opinion and non-existent information in the text	7, 12	6, 7	7, 9, 12
9	Distinguishing between fact, opinion and non-existent information in the text	6, 7, 9	6, 7	7
10	Distinguishing between fact, opinion and non-existent information in the text	6, 7, 9, 10	6, 7	7, 9, 10, 12
11	Showing sensitivity to the cohesion of the text	2, 4, 5, 8	2, 3, 4, 9, 10	4, 8

According to the information in Table 10, the intended skills match with at least one of the raters and with both of them in several cases for the items of Task 3. One striking point Table 10 suggests is that Rater 2 did not indicate a relevant skill related to Item 3. Instead, Rater 2 specified new skills on the rating scale and indicated that Item 3 requires "rereading the relevant parts" and "skimming/search reading to locate relevant information". "Rereading relevant parts" was also indicated for items 1, 4 and 8, while "skimming/search reading to locate relevant information" was also indicated for Item 8.

Table 10. Experts' Ratings Related to Task 3.

Item	Intended purpose of the item	Corresponding skills on the rating scale	Rater 1	Rater 2
1	Making inferences and drawing accurate conclusions based on explicit information from the text	6, 7, 12	6, 7, 9	6
2	Determining the meaning of idiomatic expressions from the context	6, 10	10, 11	6
3	Making inferences and drawing conclusions based on explicit information from the text	12	10, 11, 12	
4	Skimming the text and identifying implicit details from the text	7, 12	6, 7, 9	7, 12
5	Scanning the text and identifying explicit details from the text	6, 7	6, 7, 9	7
6	Identifying the author's purpose based on the explicit and implicit information from the text	1, 2, 3, 4	12	1, 2, 4, 8
7	Showing sensitivity to the cohesion of the text	8	8	8
8	Recognizing the significant points of the text and summarizing the text by identifying main ideas, themes, details or procedures	2, 3, 5	8	1, 2, 3, 5

Rereading is a strategy that can be employed while doing careful reading, and search reading can be employed to locate the related part of the text. Therefore, along with the intended skill "Making inferences and drawing conclusions based on explicit and implicit information from the text", Item 3 may also require the skill and the strategy indicated by Rater 2.

Table 11 suggests that Task 4 also includes skills that were not intended, and intended skills that were not agreed by the experts. Especially skills 9 and 10 seem to be repeated by the raters for most of the items. Thus, "Deducing meaning from context" can be relevant to Item 3, Item 4, and Item 6. Furthermore, although it is not the central aim of the items, "Distinguishing fact from opinion" might be a necessary skill on Item 3, Item 5, Item 7, and Item 8.

Table 11. Experts' Ratings Related to Task 4.

Item	Intended purpose of the item	Corresponding skills on the rating scale	Rater 1	Rater 2
1	Making inferences and drawing accurate conclusions based on explicit information from the text	6, 7, 12	6, 7	7
2	Identifying explicit details from the text	6, 7, 10	6	10, 12
3	Identifying explicit and implicit details from the text	6, 7, 12	6, 7, 9, 10	7, 12
4	Identifying explicit and implicit details from the text	6, 7, 12	6, 7	10, 11, 12
5	Identifying implicit details from the text	7, 11, 12	6, 7, 11	7, 9, 11, 12
6	Making inferences and drawing conclusions based on explicit and implicit information from the text	7, 12	12	7, 10, 12
7	Making inferences and draw conclusions based on explicit and implicit information from the text	7, 11, 12	11	7, 9, 11
8	Identifying the author's purpose based on the explicit and implicit information from the text	2, 4, 11	2	7, 9, 11, 12

Table 12 shows that three of the intended skills consistently overlap with the ratings of Rater 1 while skill 5 consistently differs from the intended skills. Rater 2, on the other hand, indicated that skill 10 might be relevant to Item 3, Item 4, and Item 5. Table 12 also shows that Rater 2 did not think Item 1, Item 6, and Item 7 as relevant to any of the skills on the list. She explained that these items, and also Item 2, can be answered with simple background knowledge by a C2 level test taker, and thus test takers may not need any other reading skills to answer these items. She also suggested that Item 3 and Item 4 may require reading carefully at sentence level.

Table 12. Experts' Ratings Related to Task 5.

Item	Intended purpose of the item	Corresponding skills on the rating scale	Rater 1	Rater 2
1	Identifying specific information from a specialized source	6, 7, 13	5, 6, 7, 13	
2	Identifying specific information from a specialized source	6, 7, 13	5, 6, 7, 13	6
3	Identifying specific information from a specialized source	6, 7, 13	5, 6, 7, 13	10
4	Identifying specific information from a specialized source	6, 7, 13	5, 6, 7, 13	6, 7, 10
5	Identifying specific information from a specialized source	6, 7, 13	5, 6, 7, 13	10, 11
6	Identifying specific information from a specialized source	6, 7, 13	5, 6, 7, 13	
7	Identifying specific information from a specialized source	6, 7, 13	5, 6, 7, 13	

Table 13 summarizes the total number of occurrences of the skills that were aimed, indicated by Rater 1 and indicated by Rater 2. Although the columns generally tend to resemble to each other, there are also discrepancies.

Table 13. Number of Occurrences of Aimed and Indicated Skills.

Skills	Aimed	Indicated by Rater 1	Indicated by Rater 2
Skimming for overall gist	2	2	2
Demonstrating understanding of text as a whole	5	4	3
Identifying topic of text	3	2	2
Identifying function of text	4	3	3
Distinguishing main points of text from subsidiary ones	2	9	1
Retrieving specific information by scanning text	25	28	4
Locating and selecting relevant factual information to perform task	32	29	24
Demonstrating understanding of how text structure works	2	2	3
Distinguishing fact from opinion	7	5	9
Deducing meaning from context	5	6	9
Interpreting text for author’s attitude and style	4	4	6
Making inferences from information given in the text	11	4	13
Making use of clues such as subtitles, illustrations	7	7	0

Tables 14 – 18 show how much raters and intended corresponding skills overlap in terms of the skills that each item is thought to measure. They show the number of

occurrences where there are overlaps and where there are no overlaps. Corresponding percentages of the number of occurrences are provided in parenthesis. Column five indicates the total number of overlaps for a specific item and its percentage.

Table 14. Overlap between Intended Skills and Expert Judgments (Task 1).

Task	Item	Overlap with 2 raters	Overlap with one rater	Total num. of overlaps	No overlap
1	1	1 (20)	1 (20)	2 (40)	3 (60)
	2	1 (20)	1 (20)	2 (40)	3 (60)
	3	1 (33.3)	1 (33.3)	2 (66.6)	1 (33.3)
	4	1 (33.3)	1 (33.3)	2 (66.6)	1 (33.3)
	5	1 (33.3)	1 (33.3)	2 (66.6)	1 (33.3)
	6	2 (33)	2 (33)	4 (66)	2 (33)

Table 15. Overlap between Intended Skills and Expert Judgments (Task 2).

Task	Item	Overlap with 2 raters	Overlap with one rater	Total num. of overlaps	No overlap
2	1	1 (12.5)	1 (12.5)	2 (25)	6 (75)
	2	1 (50)	1 (50)	2 (100)	0 (0)
	3	1 (50)	0 (0)	1 (50)	1 (50)
	4	2 (40)	2 (40)	4 (60)	1 (20)
	5	1 (20)	3 (60)	4 (80)	1 (20)
	6	1 (33.3)	2 (66.6)	3 (100)	0 (0)
	7	1 (33.3)	1 (33.3)	2 (66.6)	1 (33.3)
	8	1 (25)	1 (25)	2 (50)	2 (50)
	9	1 (33.3)	1 (33.3)	2 (66.6)	1 (33.3)
	10	1 (20)	3 (60)	4 (80)	1 (20)
	11	1 (14.3)	2 (28.6)	3 (42.9)	4 (57.1)

Table 16. Overlap between Intended Skills and Expert Judgments (Task 3).

Task	Item	Overlap with 2 raters	Overlap with one rater	Total num. of overlaps	No overlap
3	1	1 (25)	1 (25)	2 (50)	2 (50)
	2	0 (0)	2 (66.6)	2 (66.6)	1 (33.3)
	3	0 (0)	1 (33.3)	1 (33.3)	2 (66.6)
	4	1 (25)	1 (25)	2 (50)	2 (50)
	5	1 (33.3)	1 (33.3)	2 (66.6)	1 (33.3)
	6	0 (0)	3 (50)	3 (50)	3 (50)
	7	1 (100)	0 (0)	1 (100)	0 (0)
	8	0 (0)	3 (60)	3 (60)	2 (40)

Table 17. Overlap between Intended Skills and Expert Judgments (Task 4).

Task	Item	Overlap with 2 raters	Overlap with one rater	Total num. of overlaps	No overlap
4	1	1 (33.3)	1 (33.3)	2 (66.6)	1 (33.3)
	2	0 (0)	2 (50)	2 (50)	2 (50)
	3	1 (20)	2 (40)	3 (60)	2 (40)
	4	0 (0)	3 (60)	3 (60)	2 (40)
	5	2 (40)	1 (20)	3 (60)	2 (40)
	6	1 (33.3)	1 (33.3)	2 (66.6)	1 (33.3)
	7	1 (25)	1 (25)	2 (50)	2 (50)
	8	0 (0)	2 (33.3)	2 (33.3)	4 (66.6)

Table 18. Overlap between Intended Skills and Expert Judgments (Task 5).

Task	Item	Overlap with 2 raters	Overlap with one rater	Total num. of overlaps	No overlap
5	1	0 (0)	3 (75)	3 (75)	1 (25)
	2	1 (25)	2 (50)	3 (75)	1 (25)
	3	0 (0)	3 (60)	3 (60)	2 (50)
	4	2 (40)	1 (20)	3 (60)	2 (40)
	5	0 (0)	3 (50)	3 (50)	3 (50)
	6	0 (0)	3 (75)	3 (75)	1 (25)
	7	0 (0)	3 (75)	3 (75)	1 (25)

Tables 14 - 18 show that the percentage of total number of overlaps is usually higher than the number of no overlaps. It is also worth to note that there are no cases where an unintended skill is identified as relevant to a specific item by both of the raters. Therefore, any overlap that has been discussed here necessarily means overlap with the intended skills. The percentages in the tables above are item based, so in order to see the overall picture, the percentage of overlaps that are task based are calculated as the average of percentages related to individual items and presented in Table 19 below. Table 19 also shows the percentage of agreement between the two raters.

Table 19. Percentage of Overlaps.

Task	Raters' agreement with the intended skills (%)	Overlap between the two raters (%)
1	57.7	29.7
2	65.5	28.6
3	59.5	14.8
4	55.9	19.3
5	67.1	9.4

These percentages related to items show how similar the expert judgments and intended skills are. They also show that the agreement between the two raters is generally quite low. The extent of agreement of the experts with the aimed skills is not very large, but it suggests that the content of the tasks tend to comply with the expert judgments. However, it is always possible to achieve higher levels of agreement on the skills being measured; therefore, the suggestions and ratings of the experts were taken into account and the suggested revisions were done on the tasks. The suggested revisions are believed to increase the content validity of these five tasks hereafter.

4.2 Differences between proficiency groups

In order to determine whether the tasks could distinguish between higher- (TKF 315/317) and lower-proficiency (TKF 211) participants, an independent samples *t*-test was conducted on the scores from each task. The Levene's Test for Equality of Variances indicated that the two groups had equal variances on each of the five tasks ($p > .05$ for all tasks). The results are presented in Table 20 below.

Table 20. Results of Independent Samples t-tests.

	TKF 211		TKF 315/317		Highest		
	M	SD	M	SD	possible score	t	Cohen's d
Task 1	1.33	1.72	3.00	1.79	6	2.64*	.95
Task 2	2.87	3.04	6.88	2.94	11	-3.73**	1.34
Task 3	2.00	2.04	3.76	1.92	8	-2.47*	.89
Task 4	0.71	1.07	3.12	2.26	8	-3.65**	1.36
Task 5	2.53	2.36	3.81	1.9	7	-1.67	

Note. * $p < .05$, ** $p < .001$.

The descriptive statistics indicate that the TKF 315/317 group ($N = 16$) had substantially higher means than the TKF 211 group ($N = 15$) on all tasks except for Task 5. The *t*-test results show that differences between the groups were statistically

significant with large effect sizes for the first four tasks. However, no significant group difference was observed on Task 5.

The information extracted from the independent samples *t*-test analyses presents evidence for the concurrent validity of the five reading tasks in question. The results were compared to the criterion which is different levels of Turkish language proficiency anticipated depending on the participants' reports and the classes they take. Depending on the criterion, it was projected that those in TKF 315/317 classes had higher levels of proficiency than those in TKF 211 classes. The scores of the participants on Task 1, Task 2, Task 3, and Task 4 proved that these four tasks efficiently discriminated between higher and lower levels of reading proficiency. In other words, the results related to these four tasks complied with the criterion. However, the statistics related to Task 5 are not in the same line since it was found that the two groups were not discriminated efficiently on Task 5. Therefore, Task 5 was found problematic and it is wise to exclude this task from the test. The information presented in this part can be interpreted as evidence for criterion-related validity which is considered a subcategory of construct validity.

4.3 Item analysis and distractor efficiency analysis

The purpose of this section is to explore how efficiently individual items and the distractors of each item function within a task. The analyses carried out in this section investigate item difficulty, item discriminability, distractor efficiency, and Cronbach alpha values, which have the potential to improve the current version of the test.

Based on the test takers' performance on the tasks, an item analysis and a distractor efficiency analysis were carried out for each task. The tables 22 - 31

summarize the information from these analyses. A summary of score characteristics is presented in Table 21.

Table 21. Summary of Score Characteristics.

Task	Mean	SD	Minimum score	Maximum score	Maximum possible score	Cronbach's alpha
1 (B1)	2.19	1.92	0	6	6	.774
2 (B2)	4.93	3.57	0	11	11	.871
3 (C1)	2.96	2.13	0	7	8	.685
4 (C2)	3.19	2.19	0	7	8	.779
5 (C2)	2.03	2.16	0	7	7	.778

Table 21 presents the mean scores, standard deviations and Cronbach alpha values for each task. Looking at the minimum scores, maximum scores and standard deviation values, we can say that the scores of the participants are quite widespread. The mean scores for each task show that the participants generally underperformed because they scored below 50% success on average. Table 22 provides the item statistics regarding Task 1.

Table 22. Item Statistics for Reading Task 1.

Intended level	Item	Difficulty indices	Item-total correlation	Cronbach's alpha if item deleted	Cronbach's alpha for the task
B1	1	.48	.41	.770	.774
	2	.16	.57	.733	
	3	.58	.49	.749	
	4	.23	.58	.726	
	5	.39	.46	.755	
	6	.35	.64	.707	

The difficulty indices of the items in Task 1 suggest that most of the items are within the acceptable .20 - .80 interval. However, Item 2 has a value below .20, which means that the test takers found this item relatively difficult. The reason might be the fact that the test takers from TKF 211 classes, who consist of nearly the half of the participants, have generally a lower Turkish proficiency level than the other half; therefore, their underperformance could be the reason of difficulty values below .20. Despite the high item-total correlation of this item, it needs to be replaced by an easier item or revised. Although the number of items and participants are limited,

Table 22 shows that the alpha coefficients for Task 1 are quite consistent and form a basis for high internal reliability.

The distractor efficiency analysis provided in Table 23 shows the percentage of test takers who chose each of the options. The numbers showing percentage of the test takers who chose the correct answers are indicated with an asterisk.

Table 23. Distractor Efficiency Analysis for Task 1.

Options	Item 1 %	Item 2 %	Item 3 %	Item 4 %	Item 5 %	Item 6 %
A	3.22	32.25	58.06*	22.58*	6.45	3.22
B	9.67	16.12*	9.67	25.80	3.22	35.48*
C	9.67	12.9	6.45	6.45	38.7*	9.67
D	48.38*	16.12	0	12.90	29.03	19.35
Unanswered	29.03	22.58	25.80	32.25	22.58	32.25

Table 23 indicates that most of the test takers chose option A for Item 2, while the correct answer is option B. This suggests that option A is a very strong distractor, which is probably one of the reasons of this item's item difficulty value being too low. Depending on the information in Table 23, Option A of Item 1 should be revised. The fourth option of Item 3 was not chosen by any of the test takers, which means that it did not function well as a distractor. The reason might be the limited number of test takers; however, the option should be revised because the test takers may have found it implausible. Table 23 shows that all the other distractors drew some responses, and functioned well. One striking point in this table and the other distractor analysis tables is the percentage of unanswered questions. The reason can be the fact that the test takers, especially those who were in TKF 211 classes, found the tasks quite difficult and left the items blank when they felt unsure about the answer.

Table 24 indicates that most of the items of Task 2 were found to have acceptable difficulty values which are close to the medium .50 value. However, Item 11 looks problematic because its item-total correlation value is below .20. Item 11 is

Table 24. Item Statistics for Reading Task 2.

Intended level	Item	Difficulty indices	Item-total correlation	Cronbach's alpha if item deleted	Cronbach's alpha for the task
	1	.45	.57	.860	
	2	.26	.63	.856	
	3	.45	.59	.858	
	4	.32	.55	.861	
	5	.35	.44	.868	
B2	6	.58	.60	.857	.871
	7	.68	.71	.851	
	8	.55	.69	.852	
	9	.42	.76	.846	
	10	.45	.61	.857	
	11	.42	.18	.886	

a multiple-choice item with four options and the type of this item is different from the previous 10 items, so this can be the reason that some high-achieving test takers got confused and went for the wrong option. The item has an acceptable item difficulty value, but its low item-total correlation value means that this item should be omitted from the task because its dissimilarity to previous items creates confusion. The item-total correlation values for other items are quite high, which indicates that the items efficiently discriminated between high-achieving and low-achieving test takers. Furthermore, the consistent and high alpha coefficients mean that the task has high internal reliability.

Table 25 shows that all of the distractors drew some answers and worked well. Item 11 is not presented in Table 25 because its format is different from the previous 10 items. Item 11 has four options which respectively drew the following percentage of responses: option A – 3,22%, B – 25,8%, C – 41,93%, D – 12,9%, and 16,12% of the test takers didn't provide a response.

Table 25. Distractor Analyses for Task 2.

Options	Item 1 %	Item 2 %	Item 3 %	Item 4 %	Item 5 %	Item 6 %	Item 7 %	Item 8 %	Item 9 %	Item 10 %
F	45.16*	22.58	6.45	32.25*	35.48*	6.45	9.67	9.67	12.9	45.16*
G	12.90	25.80	12.9	12.90	6.45	58.06*	67.74*	12.9	41.93*	25.8
Y	19.35	25.80*	45.16*	19.35	16.12	12.90	3.22	54.83*	16.12	3.22
Unans.	22.58	25.80	35.48	35.48	41.93	22.58	19.35	22.58	29.03	25.80

Note. F = Writer's personal opinion, G = Factual information, Y = Not stated in the text, Unans. = Unaswered.

Although most of the test takers provided the correct answer for Item 11, which is Option C, this item should be omitted because of its item-total correlation value.

Table 26 shows that all of the items have difficulty values within the acceptable interval. Since the level of the task is quite demanding, it is thought to be normal that the test takers found a few items difficult. Table 26 also shows that Item 4 and Item 6 have item-total correlation values lower than .20. The alpha values when the item is deleted also suggest that these two items do not contribute much to the internal reliability of the task because the alpha coefficient increases if one of these two items is deleted. Both Item 4 and Item 6 require overall understanding of the text rather than locating specific bits of information, so the test takers might have found it quite hard to handle the majority of the information in the text because they couldn't locate any one-to-one information match between the local parts of the text and the items. It can be deduced from Table 26 that Item 4 and Item 6 need revision because they do not discriminate well and they influence the internal reliability of the task negatively.

Table 26. Item Statistics for Reading Task 3.

Intended level	Item	Difficulty indices	Item-total correlation	Cronbach's alpha if item deleted	Cronbach's alpha for the task
C1	1	.45	.41	.646	.685
	2	.26	.36	.659	
	3	.48	.66	.581	
	4	.32	.19	.696	
	5	.29	.25	.683	
	6	.58	.17	.704	
	7	.29	.50	.627	
	8	.29	.50	.627	

Table 27 indicates that Option D of Item 1 and Option A of Item 6 didn't work efficiently as distractors because they didn't draw any responses. These two distractors need revision and they should be tested again to see if the new distractors draw any responses.

Table 27. Distractor Analyses for Task 3.

Options	Item 1 %	Item 2 %	Item 3 %	Item 4 %	Item 5 %	Item 6 %	Item 7 %	Item 8 %
A	45.16*	25.80	48.38*	16.12	9.67	0	29.03*	19.35
B	6.45	25.80*	3.22	29.03	12.90	58.06*	16.12	29.03*
C	29.03	9.67	3.22	6.45	29.03*	3.22	9.67	3.22
D	0	16.12	22.58	32.25*	32.25	12.90	6.45	16.12
Unanswered	19.35	22.58	22.58	16.12	16.12	25.80	38.70	32.25

Table 28 shows that Item 4 has very low item difficulty and item-total correlation values, which means that the item was very difficult for this group of participants and it did not discriminate well between high and low achieving test takers.

Table 28. Item Statistics for Reading Task 4.

Intended level	Item	Difficulty indices	Item-total correlation	Cronbach's alpha if item deleted	Cronbach's alpha for the task
C2	1	.29	.64	.727	.779
	2	.35	.53	.746	
	3	.39	.24	.799	
	4	.10	.17	.794	
	5	.19	.61	.735	
	6	.19	.61	.735	
	7	.32	.48	.757	
	8	.19	.61	.735	

Moreover, the alpha coefficient increases when Item 4 is deleted; therefore, this item should be revised or replaced. Item 5, Item 6, and Item 8 also have low difficulty values, but they are very close to the .20 margin. Considering the task's level and these items' item-total correlation values, they can be kept to be tested again. Item 3 can be revised because of its item-total correlation value, which is close to the margin, and because the alpha coefficient increases when the item is omitted.

Table 29 shows that Option B of Item 2 and Option D of Item 7 drew no responses as distractors, which means that they need revision, and they may be replaced with new distractors. Option A of Item 4 drew most of the responses while the correct answer is C. Therefore, revision of Option A in order to make it a weaker distractor is necessary. The percentage of test takers who didn't answer Item 5, Item 6, Item 7, and Item 8 explains why these items have low item difficulty values. Since

most of the participants found these items rather difficult, they probably preferred not providing an answer.

Table 29. Distractor Efficiency Analysis for Task 4.

Options	Item 1 %	Item 2 %	Item 3 %	Item 4 %	Item 5 %	Item 6 %	Item 7 %	Item 8 %
A	16.12	16.12	38.70*	45.16	12.9	9.67	32.25*	6.45
B	16.12	0	12.9	12.9	19.35*	22.58	9.67	16.12
C	29.03*	25.8	19.35	9.67*	19.35	19.35*	9.67	9.67
D	12.90	35.48*	6.45	6.45	9.67	6.45	0	19.35*
Unanswered	25.80	22.58	22.58	25.8	38.70	41.93	48.38	48.38

Table 30 suggests that most of the items have acceptable difficulty and item-total correlation values except for the item 5 which has an item difficulty value close to the .20 margin.

Table 30. Item Statistics for Reading Task 5.

Intended level	Item	Difficulty indices	Item-total correlation	Cronbach's alpha if item deleted	Cronbach's alpha for the task
C2	1	.58	.57	.736	.778
	2	.58	.53	.745	
	3	.26	.57	.738	
	4	.65	.46	.758	
	5	.19	.51	.751	
	6	.52	.49	.753	
	7	.42	.41	.769	

However, taking into account the level of the task, this item can be tested again because its item-total correlation value is fairly high. As for reliability, the high and consistent alpha coefficients show that the task has high internal reliability with this group of participants. Spread of responses regarding items of Task 5 is presented in Table 31.

Table 31. Spread of Responses for Task 5.

Options	Item 1 %	Item 2 %	Item 3 %	Item 4 %	Item 5 %	Item 6 %	Item 7 %
Dal	58.06*	6.45	3.22	0	0	6.45	9.67
Yaprak ayası	0	58.06*	3.22	3.22	3.22	12.90	9.67
Tomurcuk	12.90	3.22	25.80*	6.45	16.12	3.22	3.22
Kulakçık	3.22	0	12.90	64.51*	3.22	3.22	0
Kını	0	6.45	22.58	0%	19.35*	6.45	19.35
Damarlar	0	3.22	6.45	12.90	3.22	54.83*	0
Yaprak sapı	6.45	3.22	6.45	0	29.03	0	38.70*
Other	9.67	6.45	6.45	0	9.67	0	6.45
Unanswered	9.67	12.90	12.90	12.90	16.12	12.90	12.90

In Task 5, the participants were expected to label the parts of a leaf after reading Text 5 and there were 7 items which were the names of the leaf parts underlined in the text. In such a task, expecting each blank to draw at least one case of all possible responses is not necessary because the underlined words or phrases here are actually not intended to be distractors to each other. One informative point Table 31 suggests is that the participants provided some other answers that were not aimed at all for the Item 1, Item 2, Item 3, Item 5, and Item 7. This requires the revision of the prompt of the task because it is clear that some of the participants didn't understand that they were expected to use only the underlined words or phrases.

Based on the students' performances, mean difficulty index and mean item-total correlation of each task were also computed and are presented in Table 32.

Table 32. Mean Difficulty and Cronbach's Alpha Values.

Task	Intended Level	Mean Difficulty Index	Mean Item-total Correlation
1	B1	.36	.52
2	B2	.45	.57
3	C1	.37	.38
4	C2	.25	.49
5	C2	.46	.51

The mean difficulty values for each task in Table 32 show that the test takers generally found all tasks difficult because all difficulty values are below .50. Furthermore, the order of the difficulty indices is not in the expected way.

4.4 Correlations of reading scores with other skills

Pearson product-moment correlation coefficients were computed to assess the relationship of the reading scores with scores from listening, writing and speaking tests administered to the same group. The results are presented in Table 33. Reading scores had significant relationships with listening, writing and speaking scores. The relationship between reading and listening was stronger than the relationship

Table 33. Relationship between Sub-tests.

Measure	*N*	*Mean*	*SD*	1	2	3	4
1. Reading	62	37.13	24.27	-	.632**	.866**	.728**
2. Writing	41	54.43	23.57		-	.610**	.527*
3. Listening	42	47.10	27.97			-	.847**
4. Speaking	22	59.40	24.96				-

Note. *$p < .05$, **$p < .01$.

between reading and writing or reading and speaking. This can be attributed to the fact that reading and listening are both receptive skills. These correlations should be approached tentatively because the sample size is small and the values are higher than expected.

CHAPTER 5

DISCUSSION

The analyses carried out on the reading tasks bore some valuable information about reliability, item characteristics and validity of the tasks under scrutiny. Based on the results, the tasks were revised. The revised tasks are presented in Appendix E. The discussion regarding each research question and revisions carried out on the tasks are presented below.

RQ 1: Do the experts agree on the operations measured by the test items as specified by the test writers?

The content validity of the tasks was explored by both qualitative and quantitative data from the experts. For content validity concerns, as suggested by Alderson et al. (1995) and Bachman (2004), a rating scale was employed in order to investigate the skills measured by the items. When the intended skills by test items and the skills designated by the experts compared, it was found that they agreed 61% on average. In other words, depending on expert evaluations, tasks tended to reflect the aimed content more than they did not. Evidence regarding content relevance can be interpreted as a form of evidential basis for score meaning (Messick, 1995), because the extent of agreement on the content being measured contributes the trustworthiness of score interpretation. However, the agreement between the two experts was quite low (20% on average). Bachman (2004) warned that weak agreement between the experts is a potential problem when investigating content validity through such an approach. One potential reason of the weak agreement between the experts in the present study is the background difference between the two experts in terms of area of research. Lack of prior training about the theoretical framework or ambiguity of the rating scale might have also contributed to such weak

agreement between the experts. Therefore, the results regarding Research Question 1 can be tentative.

Assuming that the same evidence informs us about the cognitive validity of the reading tasks, the extent to which operationalization of the framework was successful can also be discussed. In terms of operationalization of reading skill in test tasks, reading skills as defined in Khalifa and Weir's (2009) model were attempted to be measured in the design stage of the tasks. For this purpose, item specifications and items based on these specifications were developed to elicit expeditious or careful reading at global or local levels. It was previously explained in Chapter 2 that the following skills were defined related to reading by Khalifa and Weir (2009) in their model. Expeditious reading at local level is either scanning or search reading. The former one involves reading selectively to find a specific figure, name, etc. while the latter one involves locating relevant information necessary to answer a question. Expeditious reading at global level is defined as skimming which involves quickly and efficiently reading the text to get an overall understanding. Careful local reading refers to focusing on a local part of the text until the basic meaning of a proposition is established. Finally, careful reading at global level involves handling the majority of information in the text, and building a macro-structure on the basis of this. In the present study, for example, expeditious local reading was operationalized with Item 5 of Task 3 while expeditious global reading was attempted with Item 6 of Task 1. Careful local reading was operationalized with items 1 and 2 of Task 2, and finally, careful global reading was operationalized with items 7 and 8 of Task 3 (see Appendix B). However, the average agreement between the intended skills and the experts' judgment was not very high (61%). The level of agreement was proximate around this average value for all the five reading tasks. Moreover, the agreement

between the raters was quite low. These values raise concerns about the extent to which the reading tasks succeeded in operationalizing the intended skills. Poor or weak agreement on the skills being measured can be indicating that the aimed reading skills are not well represented on the items. Furthermore, the experts indicated that some of the skills they rated were overlapping with each other. These issues can be investigated through a further study in which, as Alderson et al. (1995) suggest, experts are given some definitions of the underlying theory and asked to make judgments about the test in terms of construct validity. In addition, more reliable evidence could have been brought in through the analysis of the cognitive operations that the test takers used in their attempt to respond to items.

In the text selection process, texts that were potentially biased in terms of background knowledge were avoided because subject or cultural knowledge in a given topic facilitates reading comprehension (Carrel, 1987; Leeser, 2007; Sabatin, 2013). Absence of such knowledge, on the other hand, should not disadvantage a group of test takers (Alderson, 2000). When selecting texts, the concerns of grading the texts in accordance with the aimed proficiency level brought in the issue of the extent to which cultural elements were tolerable. For example, the text aiming C1 level test takers was from the area of history, and it included more cultural elements, such as lexical elements, when compared to B2 level text. Similarly, Text 4 (C2) was a literary text, and the reflection of culture on the language of the text was unavoidable. Therefore, it was observed that as the aimed proficiency level of Turkish texts increased, the cultural intrusion was more probable. Since there are no studies regarding selecting Turkish texts for the purposes of Turkish reading assessment, it is not clear to what extent cultural elements are tolerable at different proficiency groups. Regarding the appropriateness of the texts, Expert rating form

revealed some concerns related to the texts, as well as instructions and questions. The feedback from the experts signaled that some revisions needed to be done on the indicated parts of the texts because of their linguistic complexity or requiring background knowledge. There are no formulas to measure linguistic complexity or tools to measure text difficulty in Turkish. Therefore, it is hard to know whether any revision concretely succeeded in reducing linguistic complexity or changed text difficulty. However, based on the suggestions by experts, some sentences, phrases and words were simplified or replaced on Text 1, Text 2, Text 3 and Text 4, and revisions were done on questions and instructions. Especially, some embedded clauses were reduced because the experts thought that embedded clauses could be potential source of difficulty, and reduction of these clauses can eliminate the problem.

Regarding the response format, in order to avoid the reservations that there may be other plausible answers than the intended correct answer depending on test takers' subjective interpretation of the text and the items (Chen, 2010), the tasks were tried on educated native speakers before the piloting. Then, the native speakers were asked to explain why they went for the options they chose in order to check whether the items are vulnerable to their subjective ideas. This was also helpful in detecting whether options were providing clues (Haladyna et al., 2002).

RQ 2: Do the test tasks differentiate between higher and lower proficiency groups?

The tasks' discrimination efficiency on different proficiency levels was investigated through statistical analysis. The aim of this analysis was to explore external relationship of the reading tasks. As Messick (1995) suggests, external relationships of a test are criterion related evidence. Depending on the class (TKF

211 or TKF 315/317) test takers take and their instructor's report, it was assumed that the two groups are at different proficiency levels. Setting their different proficiency levels as the criterion, the scores from the reading tasks were analyzed to investigate the extent to which the reading tasks reflect the criterion. The results indicated that, except for Task 5, all of the tasks discriminated between the higher and lower proficiency test takers at a significant level. Therefore, criterion related evidence, which is one of the six aspects of construct validity (Messick, 1995), was found for four of the tasks. Evidence found for the discrimination efficiency of the tasks can be a proof that formulation of what is easy and difficulty on the tasks was justified. This formulation can include parameters such as text selection (topic, length, and syntactic complexity), task selection and response format.

RQ 3: What are the psychometric characteristics of the items for each reading task?

Research question three investigated the psychometric characteristics of individual items as well as distractor efficiency. Such evidence supports reliability (or scoring validity) of the measure (Weir, 2005). Scoring validity is important because we should be able to depend on scores (Khalifa and Weir, 2009). Any evidence that support the trustworthiness of scores is connected to construct validity (Messick, 1995). On the other hand, a reduction in scoring validity runs the risk of construct irrelevant variance (Khalifa and Weir, 2009). The difficulty and discrimination indices and internal consistency values of items are accepted among the sources of evidence for reliability (Bachman, 2004; Weir 2005). In a reading test, we can relate the concerns of scoring validity to statistical item functionality. The observations and revisions based on the results regarding research question three are as follows. The items of Task 1 had quite favorable discrimination values; however,

it was revealed that the items' level of difficulty generally do not match the expectations. Taking into account the fact that this is the first trial of the items, it was decided that difficulty values between .20 and .80 were acceptable following Bachman (2004). Task 1 was intended to be the least demanding task among the 5 tasks tested, but the order of difficulty indices disproved this. Therefore, the concerns of the experts regarding Task 1's difficulty level were justified by the difficulty indices. One reason of Task 1's being too demanding can be the contextual features of Text 1. For example, the word count of Text 1 was a little higher than Text 2 (B2). Furthermore, words-per-sentence index, although smaller than Text 2, was quite similar to that of Text 3 (C1) and Text 4 (C2). Therefore, the contextual features of Text 1 might have contributed to the difficulty of Task 1 because longer texts and sentences tend to require more cognitive load (Khalifa and Weir, 2009). In accordance with the feedback from Rater 2, Text 1 was simplified. After the revisions on Text 1, the word count dropped from 371 to 332, and words per sentence dropped from 14.3 to 10. Item 1 was revised and Item 2 was replaced with an easier item that requires expeditious reading at local level to find specific information. Previously, this item required careful reading at global level, which is predicted to be more demanding than expeditious local reading by Weir and Khalifa (2008). Since the distractor analysis showed that Option D of Item 3 did not attract any answers, the wording of the option was changed. After the revisions on the text and items of this task to adjust its level, a second trial on a similar group of participants may bear more favorable results.

In general, Task 2 had rather favorable item difficulty and discrimination values, as well as efficiently working distractors. However, a number of revisions were also done on Task 2. Depending on the feedback from Rater 2, the language of

the text related to this task was simplified using a more straightforward language. For example, the sentence “Kendisinin karizması, etkileyici hatlardan ve havalı bir duruştan değil, gayet insani, gayet sıradan ve sahici bir portre çizmesinden kaynaklanıyordu” was rewritten in the following way: “Hoffman etkileyici bir fiziğe sahip değildi. Karizması, çok insani ve çok sıradan ama bir o kadar da sahici bir portre çizmesinden geliyordu”. This kind of instances showed that use of pronouns and negation need to be taken care of besides structure and length. Moreover, the order of the items was also revised so that they follow the order of processing the text as suggested by Khalifa and Weir, (2009). The item analysis on Task 2 indicated that Item 11 had a low item-total correlation value (.18) and did not efficiently discriminate between high-achieving and low-achieving test takers. Therefore, this item was omitted from the task in order to increase the overall reliability and discriminating efficiency of Task 2.

With Rater 2’s suggestion regarding Task 3, the wording of the prompt on Item 7 was revised. The format of Item 8 was changed and the options related to this item were deleted because Rater 2 indicated her concerns that options may function as clues. Haladyna et al. (2002) advise avoiding clues in the options. Such clues can lead to construct irrelevant easiness (Messick, 1995). The revised format requires test takers to indicate the order by writing down the sentence numbers in the correct order. The item analysis on this task showed that Item 4 and Item 6 did not discriminate well and negatively influenced the internal reliability of the task. A common feature between Item 4 and Item 6 was that they both had options with similar content words. This indicates that options with similar content words can lead test takers to respond wrongly; therefore, such options could be challenging for this group of test takers. These items and their options were revised. For example, Option

B of Item 4 was revised on Rater 2's suggestion because it was a very strong distractor and drew as many answers as the correct option. Therefore, instead of the sentence "Hollanda'da lale endüstrisi Leiden Üniversitesi öncülüğünde başlamıştır", the sentence "Osmanlı'da lale endüstrisi üniversiteler öncülüğünde başlamıştır" was used in order to make it a weaker distractor. It should be also noted that since the aimed proficiency level of this task is beyond the proficiency level of participants, the .37 mean difficulty of this task supported the fact that this task is appropriate for higher proficiency levels. Therefore, the things that did not work on C1 and C2 level tasks in this study can prove to be working with higher proficiency level test takers.

Difficulty and item-total correlation values of Item 4 of Task 4 were found to be unacceptable in item analysis. A close examination of this item revealed that the reason might be related to its options. A revision was done by changing the wording of Option A of this item from "1980 yılında olmuştur" to "1980 yılının başında olmuştur" since Option A turned out to be a very strong distractor and attracted nearly half of the answers. Moreover, Option B of Item 2 and Option D of Item 7 were also revised because they did not function well as distractors. Finally, with Rater 2's suggestion, the wording of Option C and Option D of Item 3 were changed for clarity concerns. Nevertheless, the data from this group of participants are also tentative because their proficiency level may not be matching the task's requirements. Since this task was expected to be beyond this group of test takers' proficiency level, the .25 mean difficulty value of the task justified the expectations. Therefore, better data can be collected from a sample with higher proficiency level.

Task 5 was found to have item difficulty and item-total correlation values within the acceptable range. Moreover, it also had a high alpha coefficient, which is a proof of high internal reliability. However, Task 5 was found to have a lower mean

difficulty value than expected. It was concluded that the reason why Task 5 was found relatively easy compared to lower level tasks can be related to item type and contextual features. Hsueh-chao and Nation (2000) suggested that readers need to know around 98% of the vocabulary for uninterrupted reading. However, the lexical complexity of Text 5 might have been ruled out by the task's requirement because labeling parts of a leaf didn't require commanding all of the vocabulary in the text. It is also possible that the length of Text 5, which was relatively short compared to the other texts, made it less demanding. Longer texts can make more demands on both higher and lower level processing (Khalifa & Weir, 2009). Furthermore, another explanation can be the fact that Task 5 was found easier because of the genre of the reading text. Task 5 included an expository text while Task 4, the other C2 level task, included a narrative one. Depending on Zhou's (2011) study results which indicated that students tended to perform better on expository texts than narrative texts, the expository nature of Text 5 might have added to the facility of the task. CEFR descriptors indicate that individuals at this level can read and understand nearly all types of written material. With no restrictions on the type of text, a technical text was chosen for this task to eliminate background knowledge because it included specialized terminology that test takers would possibly be unfamiliar with. However, the results showed that specialized topics or lexical complexity may not guarantee the difficulty of a text. It was also observed in the text selection process that technical texts tend to have shorter sentences and usually the tense of the sentences do not change much throughout the text. In literary texts, on the other hand, sentences are much longer and tense shift is more flexible. Therefore, words per sentence can be one source of evidence for text difficulty, but it is tentative and

should be used along with other measures. Such issues should be taken care of in the text selection process.

RQ 4: Do scores on the reading test correlate with the scores obtained from listening, writing and speaking tests administered to the same group?

Correlating different components of a test bears evidence related to construct validity (Alderson et al., 1995). The different test components are expected to measure different skills; therefore, the correlations should not be too high, for example +.9. Since very high correlations between two components of a test mean that they measure essentially the same thing, moderate correlations should be expected, which means the components are moderately interrelated and each component uniquely contributes to overall language proficiency (Alderson et al., 1995). The present study found significant positive correlation coefficients: .632 between reading and writing, .866 between reading and listening, and .728 between reading and speaking. Positive but weaker relationships were also found in a large scale study carried on a standardized language test. Liao, Qu, and Morgan's (2010) data from more than 12,000 TOEIC test takers bore the following correlations: .76 between reading and listening, .57 between reading and speaking, and .61 between reading and writing. On the other hand, Wang's (2008) study on College English Test Band 4 (CET-4) did not find any significant relation between reading and writing or between reading and listening. The correlations in the present study can be interpreted as evidence for the construct validity of the tasks because the components measure related abilities. The highest correlation was between reading and listening, which was an expected result because they are both receptive skills. It was also expected that the relationship of reading tasks would be weaker with writing and speaking tasks. However, these values should be approached with caution because

the values are quite large when compared to other studies. Nevertheless, the studies we compare our results here are from language tests in English. Whether the large correlations in the present study stem from the language of the test can be investigated. Empirical findings regarding this issue can only be obtained through further research.

A further issue to note here is related to the design of the tasks. Items that require expeditious reading were limitedly employed in this study, and these items were on Task 1 and Task 3 in which careful reading was also expected. This can be one of the reasons of test takers' using their time inefficiently as observed in the administration of the tasks. For example, with Item 6 on Task 1, expeditious global reading was attempted; however, the previous items requiring careful local reading on this task might have made test takers respond to Item 6 based on the local readings they had already done. Therefore, it could be a better idea to employ items that require expeditious reading and items that require careful reading on separate texts or separate components. This can guide test takers to use their time more efficiently since texts aiming expeditious reading would be allotted less time than ones aiming careful reading (Hughes, 2003).

Depending on the study results, the reading tasks investigated here are promising in the sense that they both formed an empirical grounding for future research and they can be further developed and expanded with new tasks to form the reading component of a standard test of Turkish for foreigners. It is expected that the revisions discussed above will increase the content validity, concurrent validity and reliability - all of which are accepted as aspects of construct validity - of the tasks being tested. This can only be revealed through testing the tasks on a similar group of participants again.

From the researcher's perspective, the test development process as a whole was quite educating in the sense that planning and producing reading tasks for a reading test required meticulous effort. Experiences from the test development process and contributions of experts were very educative in teaching what route to follow and what to avoid. This study has demonstrated that selection of texts in Turkish and analysis of them should be carefully carried out. Both in the text selection process and when developing items, the framework that the test is based on should be the guide, and experts should be consulted in the process. It is also important to know that test development is a recursive process and improvement of the test is always possible in the light of new data or with new sources of data.

CHAPTER 6

CONCLUSION

The study was carried out to explore the reading tasks, which were developed to test reading ability in Turkish for foreigners, in terms of content relevance, criterion related evidence and reliability. The study also investigated the correlational relationship of the reading tasks with other language abilities assessed by tasks that were developed by other researchers. The main empirical findings are as follow:

First, the expert judgments indicated that although not very strong, the content of the tasks tended to be relevant to the aimed target domain. This finding revealed the parts that are weak or obscure in terms of the abilities aimed to be measured. Therefore, the results urged revisions on the tasks. Moreover, the results signified that operationalization of the aimed skills should be carried out with great care. Since skill operationalization is based on test writer assumptions, operationalization should be a recursive process in which data collection and expert consultancy are employed to examine how well the framework is represented on items.

Second, except for Task 5, the tasks were found to be efficiently discriminating between higher and lower proficiency levels.

Third, psychometric characteristics of individual items were generally favorable, and problematic items were revised, replaced or omitted. The findings regarding psychometric characteristics of items and revisions on the problematic parts are expected to increase the reliability of the scores.

Finally, the correlations between the reading tasks and listening, writing and speaking tasks were all significant. This indicated that the tasks regarding the four language skills measured highly related language abilities.

After the revisions on the tasks, four of the reading tasks are ready to be tried on a similar sample again. However, Task 5's unfavorable difficulty level and inefficiency in discriminating proficiency levels made the researcher to decide that this task is beyond revision and should be discarded.

In conclusion, the findings of this study presented empirical evidence regarding the indicated issues above; furthermore, they also indicated the weak parts of the tasks. Such evidence is crucial when evaluating the assessment tool and improving it. Although the weak parts were responded as an attempt to improve the tasks, they can be further developed by more trials and with expert support. Given the scarcity of research on assessing reading in Turkish as a foreign language and lack of guidance in selecting appropriate texts for a Turkish reading test, these findings can prove helpful for future development in the area. Research in this area needs to be expanded because standard assessment instruments with appropriate psychometric properties to assess Turkish language are needed.

6.1 Limitations

As it may be the case in other small scale research studies, the sample size and the number of reading tasks tested in this study were limited. A larger sample size would increase the reliability of the data in such a study.

The second limitation was regarding expert judgments. On Expert Judgment Form, judgments between the two experts were inconsistent when they marked the relevant skills related to each item. Although the correlation between judgments of the two experts was low, taking two viewpoints about the skills individual items were measuring was fruitful. Moreover, although the agreement of the experts with the aimed content of the tasks was not very weak, this finding revealed that there

could be deficiencies in the operationalization of the aimed skills. To this respect, more expert support could be needed in the development and evaluation of the tasks.

The third limitation is that because of time limitations, feedback from experts was collected at the same time when data were collected. Therefore, it was not possible to do revisions on the tasks based on the feedback from the experts before the tasks were administered. The revisions were only carried out after the administration of the tasks. Therefore, even after revisions, the tasks are not in their final version and they should be tried again as a part of a recursive process until they are proven to be working efficiently.

6.2 Suggestions for future researchers

A primary suggestion is that more reading tasks aiming different proficiency levels should be developed. Creating a pool of various reading tasks that are proved to be efficiently working with the target test takers will give the opportunity to test a sample of more various skills from the target domain (Hughes, 2003). Moreover, although the tasks under investigation mostly employed multiple choice item formats, it is better to include, as Alderson et al. (1995) suggest, different item formats as well in order to make sure the test is not biased towards a particular method.

A second suggestion for future researchers is about data collection for content relevance. The experts should be consulted both when deciding on the content to be covered and when evaluating to what extent the content was operationalized. It is also advised that the researchers inform all the experts sufficiently about the theoretical framework behind the tasks.

Since the cognitive processes that test takers go through while dealing with reading tasks in Turkish are not known, it is further suggested that a cognitive

validity analysis through eye-tracking or think-aloud procedures should be carried out to better understand the processes test takers go through while completing the tasks. Eye-tracking research has the potential to reveal what kind of order test takers follow while dealing with texts and questions, or which part of the text they mostly focus on and whether this complies with the assumptions of careful and expeditious reading at different levels. Think-aloud procedures, on the other hand, can reveal how test takers interpret the given texts as well as what kind of reading strategies they adopt to complete the tasks.

APPENDIX A

LEARNER PROFILE FORM

Last Name, First Name: _____

Sex: Female _____ Male: _____

Date of Birth: _____

Place of Birth: City: _____ Country: _____

Mother Tongue: _____

Language of Education: _____

How long have you been learning Turkish? _____

How long have you been living in Turkey? _____

Contact (Mobile phone or email address): _____

Turkish Language Proficiency

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

	A1	A2	B1	B2	C1	C2
Listening	I can recognise familiar words and very basic phrases concerning myself, my family when people speak slowly and clearly. ☐	I can understand phrases and the highest frequency vocabulary related to areas relevant to my interests (e.g. very basic personal and family information, shopping, local area, employment). I can catch the main point in short, clear, simple messages and announcements. ☐	I can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure, etc. I can understand the main point of many radio or TV programmes on current topics of personal or professional interest when the delivery is relatively slow and clear. ☐	I can understand extended speech and lectures if the topic is reasonably familiar. I can understand most TV news and current affairs programmes. I can understand the majority of films in standard dialect. ☐	I can understand extended speech even when it is not clearly structured and when relationships are implied and not signaled explicitly. I can understand television programmes and films without too much effort. ☐	I have no difficulty in understanding any kind of spoken language, whether live or broadcast, even when delivered at fast native speed. ☐
Reading	I can understand familiar names, words and very simple sentences. ☐	I can read very short, simple texts. I can find specific information in simple everyday material such as advertisements, prospectuses, menus and timetables and I can understand short simple personal letters. ☐	I can understand texts that consist mainly of high frequency everyday or job related language. I can understand the description of events, feelings and wishes in personal letters. ☐	I can read articles and reports concerned with contemporary complex problems. I can understand contemporary literary prose. ☐	I can understand long and complex factual and literary texts. I can understand specialized articles and longer technical instructions, even when they do not relate to my field. ☐	I can read with ease virtually all forms of the written language, including abstract, structurally or linguistically complex texts, factual and literary texts, such as manuals, articles and literary works. ☐

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

APPENDIX B

READING TASKS BEFORE REVISION

OKUMA-ANLAMA

Bölüm 1 (6 Soru)

Aşağıdaki metni okuyarak her soru için doğru yanıtı işaretleyiniz.

Köpekbalıklarının Soyu Tükeniyor

Yeni yapılan bir araştırmaya göre okyanuslardaki köpekbalıklarının yarısından çoğu, soylarının tükenmesi tehlikesiyle karşı karşıya. Uluslararası Doğa Koruma Birliği'nden (IUCN) uzmanlar, 11 köpekbalığı türünün yüksek risk listesinde olduğunu söylüyorlar. Üstelik 5 türün daha bu listeye girme olasılığı var. Köpekbalıkları çok yavaş ürüyor ve bu nedenle aşırı avlanmadan çok etkileniyorlar. Bilim insanları, köpekbalıkları için küresel avlanma sınırlamaları getirilmesini, yüzgeçleri için avlanmalarına son verilmesini ve hata sonucu yakalanmalarını en aza indirecek önlemlerin alınmasını istiyor.

IUCN Köpekbalığı Uzman Grubu'ndan Sonja Fordham, "Çok değişik özellikleri olan köpekbalığı türleri var. Bu nedenle insanlar köpekbalıklarının aşırı avlanmaya karşı dirençli olduğunu sanıyor ama bu doğru değil" diyor. Fordham ayrıca şunları da ekliyor "Aslında köpekbalıkları için uluslararası yakalama sınırları getirilmediğinden, giderek endişe duyulan türler arasına giriyorlar. Okyanuslarda balıkçılığın yoğun yapıldığı alanlar var ve oralardaki köpekbalıkları çoğunlukla korunmasız."

Yeni Tehditler

Bilim insanları, okyanusların üst katmanlarında yüzen 21 tür köpekbalığına yönelik araştırmalardan elde edilen verileri değerlendirdi. Bu değerlendirmeye göre, 21 tür içinden birinin, dev şeytan vatozlarının, soyu tükenmek üzere, 10 türün de tükenme tehlikesi var. Geri kalan beşindeyse azalma oranı çok ciddi olmadığı için yalnızca 'tükenmeye yakın' olarak tanımlandı.

Sınıflandırmalar, popülasyonda geçmişte görülen azalmalara ve gelecekteki olası azalmalara dayanan bir dizi ölçüte göre yapılıyor. Örneğin, 10 yıl içinde nüfusu %50 oranında düşen bir tür, soyu tehlikede olarak tanımlanıyor.

Yüzgeç Kesimi

Köpekbalıklarına yönelik en önemli tehdit, bilerek ya da yanlışlıkla yapılan avlanma. Fordham, "Köpekbalıkları, önceleri kılıçbalıklarını avlayan gemiler tarafından yanlışlıkla yakalanıyordu. Ama şimdi sayıları azaldıkça köpekbalıkları, balıkçıların özellikle hedefi oluyor. Bazı türler yüzgeçleri ve eti için, bazı türler de yalnızca yüzgeçleri için avlanıyor." diyor.

Uluslararası sularda köpekbalığı avcılığını düzenleyen birçok organizasyon var. Bu organizasyonlar, yüzgeçleri için köpekbalığı avcılığını sınırlamak üzere çeşitli önlemler aldı ama her biri değişik standartlar uyguluyor. Bu da avcıların bu düzenlemelerdeki yasal açıkları bulup onları kolayca ihlal etmesine olanak tanıyor. Koruma grupları, Doğu Asya ülkelerinin ekonomilerindeki büyümenin, köpekbalıklarının yüzgeçleri için avlanmasını arttırdığını söylüyor. Raporun baş yazarı, Simon Fraser Üniversitesi'nden Nicholas Dulvy "Balıkçılık yetkililerine ve konuyla ilgili bölgesel, ulusal ve uluslararası yetkililere bu durumu düzeltmek için büyük yükümlülük düşüyor. Aslında durum böyle olmak zorunda değil. Güçlü bir halk desteği ve politik kararlılıkla bu düşüş tersine çevrilebilir." diyor.

Rapor, Bonn'da yapılan Biyolojik Çeşitlilik Konvensiyonu'nda sunuldu. Rapor aynı zamanda Sucul Yaşamın Korunması: Deniz ve Tatlı Su Ekosistemleri adlı dergide bu yılın sonunda yayımlanacak. Dergide, IUCN'nin Tehdit Altındaki Türler Kırmızı Listesi'nin yeni risk değerlendirmesi de olacak.

Adapted from
Korkut Demirbaş
BiLiMveTEKNiK 4 Temmuz 2008

1) Bilim insanlarına göre aşağıdakilerden hangisi, soyları tehlikede olan köpekbalıkları için alınması gereken önlemlerden biri değildir?

A) Uluslararası avlanmanın kontrol edilmesi
B) Yüzgeçleri için avlanmasının yasaklanması
C) Kazara yakalamanın azaltılması
D) Daha fazla üremeleri için uygun şartların yaratılması

2) Metne göre göre aşağıdakilerden hangisi doğrudur?

A) Toplam 16 tür köpekbalığı yüksek yok olma riski altındadır.
B) Yok olma tehlikesi olan köpekbalığı türlerinin sayısı olmayanlardan daha fazladır.
C) Köpekbalıklarının üreme hızı yok olma sebeplerinden biri değildir.
D) Okyanuslarda avlanan balıkçılar giderek daha az köpekbalığı yakalamaktadır.

3) Eskiden balıkçıların köpekbalığı yakalamasıyla ilgili aşağıdakilerden hangisi doğrudur?

A) Yanlışlıkla yapılırdı.
B) Onları çok uğraştırırdı.
C) Sadece etleri için yapılırdı.
D) Sadece kılıçbalığı azaldığında olurdu.

4) 6. paragrafa göre uluslararası sularda köpekbalığı avcılığını sınırlama çabaları neden işe yaramıyor?

A) Farklı uygulamalar yasal açıklar yarattığı için.
B) Doğu Asya'da köpekbalığı yüzgeçleri popüler olduğu için.
C) Bazı köpekbalığı türleri hala risk grubunda görülmediği için.
D) Hükümetler bu konuyu yeterince ciddiye almadığı için.

5) Nicholas Dulvy'e göre aşağıdakilerden hangisi köpekbalığı türlerini yok olma tehlikesinden kurtarmak için gereklidir?

A) Köpekbalığına dayalı ekonomilerin küçültülmesi
B) Balıkçılara eğitimler verilmesi
C) Güçlü bir halk desteği
D) Yetkililerin bu konudaki politik görüşlerini değiştirmesi

6) Bu yazıda yazarın genel amacı aşağıdakilerden hangisidir?

A) Köpekbalıklarına karşı olan çevresel duyarsızlığı gözler önüne sermek
B) Soyu tehlikede olan köpekbalıkları konusunda farkındalık yaratmak
C) Özellikle hangi tür köpekbalıklarının risk altında olduğunu açıklamak
D) Köpekbalığı sayılarındaki azalmanın çeşitli sebeplerini incelemek

Bölüm 2 (11 Soru)

Aşağıdaki metni okuyarak her soru için doğru cevabı işaretleyiniz.

Hoffman'a Veda

1 Oynadığı her filme kendi hissiyatını katan, filmde kendi varlığını hissettiren oyunculardan olan Philip Seymour Hoffman artık aramızda değil. 2 Şubat günü evinde, kolunda bir şırıngayla ölü bulundu. Malum, aşırı doz. Bu alçak gönüllü ve mağdur figürün bu şekilde, henüz kırk altı yaşındayken ölmesi üzücü fakat şaşırtıcı değil. Canlandırdığı hafiften kaybeden, hayatın kenarında duran, trajik ve krizli karakterlere uygun bir son. Yani bir bakıma Hoffman, yaşadığı gibi oynamış, oynadığı gibi de hayata veda etmiş oldu.

2 Hoffman gerçek bir 'yıldız'dı ama 'yıldız' sistemine uygun bir tavıra ve duruşa sahip değildi. Aslında 'silik' olabilecek olağan hatları, hafif tombulluğu, hantallığı ve döküklüğüyle sorunlu 'yan komşu' tiplemesine ve daha doğrusu genel olarak yan rollere uygundu. Ama yan rollerde belirdiğinde bile insanın aklına bir çengel gibi takılıyor, başrolden 'rol' çalıyordu. Kendisinin karizması, etkileyici hatlardan ve havalı bir duruştan değil, gayet insani, gayet sıradan ve sahici bir portre çizmesinden kaynaklanıyordu. İnsanda içten samimi bir sevgi yaratmasının nedeni de buydu herhalde.

3 Bu gerçeklik sayesinde 'iyi insan' rollerini de 'kötücül' karakterleri de müthiş bir sahicilikle canlandırıyordu. Aslında tıpkı yan rol/baş rol ayrımını ortadan kaldırdığı gibi iyi/kötü ayrımını da ortadan kaldırıyordu. Belki de bu yüzden, bir kategoriye kolayca yerleştirilip kenara koyulamadığı için filmlerdeki varlığı seyircilerin hep aklında kalıyordu.

4 Paul Thomas Anderson'ın neredeyse bütün filmlerinde kadrolu oyuncu olan Hoffman'ın oynadığı son Anderson filminin, sonunu hazırlayan bir olaya vesile olması son derece ironik. Corinne Van Vliet'in yazısına göre, Hoffman en son 1989'da veda ettiği madde bağımlılığına, *The Master*'ın 2012'de yapılan galasında kendisine ikram edilen bir kadeh içkiyle geri dönmüş. Daha doğrusu şöyle: tam yirmi üç yıldır alkolden ve uyuşturuculardan uzak duran Hoffman ilk kez bu galada alkolü tekrar bünyesine sokmuş ve işte o alkol damlası Hoffman'ı tekrar bağımlılığa götüren yolun başlangıcı olmuş. Hızla yenilenen eski bağımlılık Hoffman'ı iki seneden kısa bir süre içinde aşırı dozdan ölüme götürmüş. Bu son filmin adının 'Usta' olması da Hoffman'a atfedilebilecek bir unvan olması sebebiyle hoş (ve acı) bir ironi içeriyor. Trajik ve erken ölümü de filmlerle geçen hayatına bir film sahnesi gibi eklenmiş oldu. Huzur içinde yatsın.

Ahmet Ergenç
Altyazı Dergisi

A. Aşağıdaki cümleleri okuduğunuz metne göre değerlendiriniz. Her cümle için F, G ve Y'den birini işaretleyiniz.

F: Eğer verilen cümle yazarın kişisel fikriyse
G: Eğer verilen cümle nesnel bir gerçekten bahsediyorsa
Y: Eğer verilen cümle bu yazıda yer almayan bir ifadeyse

1. Hoffman'ın ölümü, canlandırdığı karakterlere uygun bir son oldu.	**F**	**G**	**Y**
2. Hoffman neşeli, hayat dolu karakterleri canlandırmak istemiştir.	**F**	**G**	**Y**
3. Hoffman'ın bağımlılığı ailevi nedenlerden kaynaklanıyordu.	**F**	**G**	**Y**
4. Hoffman oynadığı rollere samimiyet katabiliyor, bu da onu etkileyici bir oyuncu yapıyordu.	**F**	**G**	**Y**
5. Oynadığı rollerden dolayı Hoffman'ı hemen etiketlemek zor olduğu için seyircinin aklında kalıyordu.	**F**	**G**	**Y**
6. Hoffman, yönetmen Anderson'un filmlerinin çoğunda rol alıyordu.	**F**	**G**	**Y**
7. "The Master" filminin galasına kadar Hoffman, çok uzun bir süredir alkol kullanmıyordu.	**F**	**G**	**Y**
8. Hoffman bağımlılıktan kurtulmak için 2012 yılına kadar tedavi görmüştür.	**F**	**G**	**Y**
9. Hoffman'ın ölüm sebebi uyuşturucudur.	**F**	**G**	**Y**
10. Hoffman yan rollere kattığı hakikilikle bir yıldızdı.	**F**	**G**	**Y**

B. Aşağıdaki cümleler hangi paragrafın sonuna gelebilir?

Hoffman'ın oynadığı bazı karakterlere bakacak olursak bunu açıkça görebiliriz. Örneğin "Big Lebowski" filmindeki Brandt, "Red Dragon" filmindeki Freddy Lounds ve "Hunger Games" filmlerindeki Plutarch Heavensbee karakterlerini kolayca sınıflandırmak hiç kolay değil ve tabii unutmak da.

A) 1 B) 2 C) 3 D) 4

Bölüm 3 (8 Soru)

Aşağıdaki metni okuyarak her soru için doğru cevabı işaretleyiniz.

Lale

Lale, zambakgiller ailesinden, yaprakları uzun ve mızraksı, çiçekleri kadeh biçiminde, türlü renkte, alacalı bir süs bitkisidir. Laleler şimdilerde, birbirinden farklı renkleri, zerafetleriyle İstanbul'un cadde ve parklarını süslüyorlar. Özellikle son zamanlarda tekrar tanıştığımız lale, hiçbir çiçekte olmadığı kadar insanları etkisi altına almış ve renkli bir öykünün baş kahramanı olmuştu.

İnsanları hastalık derecesinde kendisine tutkun eden lalelerin öyküsü Osmanlılarla başlamıştır. İstanbul'un Fethi'nden sonra, Fatih'in emri ile yeniden düzenlenen bahçeler (parklar) lâlelerle süslenmiştir. Zaten Fatih Sultan Mehmet bir bahçıvandı. Boş vakitlerinin çoğunu bunun için harcar ve bundan büyük bir haz duyardı. Boş zamanlarda Topkapı ve diğer sarayların bahçelerinde çalışmaktan da büyük zevk alırdı.

Kanuni devrinde de, lale türleri geliştirip çoğaltılmıştır. 16. yüzyılda İstanbul'da yayılıp, oradan da Avrupa'ya kadar sıçramıştı. Muhteşem Süleyman döneminde küçük ve kısa boylu, çiçeği badem biçiminde, uçları ince ve sivri, İstanbul'a özgü lâle çeşitleri yetiştirilmiş, bilmeden de olsa lale çılgınlığı böylece başlatılmıştı. İnce mi ince, uzun mu uzun laleler öylesine nazlı görünüyorlardı ki, bir kere gören etkisinden kurtulamıyor, böylesi bir güzelliğe sahip olabilmek için akla gelmez çılgınlıklar yapılıyordu.

Kimler yoktu ki, bu çılgınların arasında. Veziri, sadrazamı lalelerle ilmî olarak ilgileniyor, nadide bir lâle soğanına servet ödendiğini bilenler ise, yeni bir çeşit bulmanın hayaliyle bahçelerinde gizli gizli deneyler yapıyordu. Ardı ardına yeni lale çeşitleri çıkmış, bunlara da görüntülerinin güzelliğine yakışır, pırıltılı adlar takılmıştı. Kimileri laleye "gönül yakan" adını vermeyi uygun görmüş, kimisi şans getireceğine inanarak "talih yıldızı" demiş, bazıları da hissettiklerine tercüman olması için, "sevinç ışığı" adını vermişti çiçeğine.

Ancak laleler sadece güzelliklerle anılmıyor, kimi zaman da polisiye olaylarla gündeme geliyordu. Örneğin, lale çılgınlığının en üst boyutlara ulaştığı Lale Devri'nde, Taç-ı Kayser adı verileren nadir bir lale türü Çırağan Sarayı'nın bahçesine ekilmiş ancak çalınmıştı. Damat İbrahim Paşa, işin peşini bırakmamış. İlk şüpheliler lale tutkunları olduğu için, gizlice bahçeleri aranmış. Nafile ki tüm çabalar boşa çıkmıştı.

Avusturya-Macaristan İmparatorluğu'nun Kanuni Sultan Süleyman nezdindeki büyükelçisi Ogier Ghislain de Busbeck'in 1554 yılında geldiği İstanbul'dan Avusturya'da yaşayan dostu Carolus Clusius'a lale soğanları gönderdiği sanılmaktadır. Daha sonra Hollanda'ya giderek Leiden Üniversitesi'nde göreve başlayan Clusius, bu ülkelerde laleyi ilk yetiştiren ve lâle endüstrisini kuran kişi olarak bilinmektedir. Avrupa'ya giden lale, özellikle Hollanda ve Almanya'da aranan bir meta haline gelmişti ve bir lale soğanına bütün servetini yatıranlar vardı. Hollanda, günümüze kadar gelen lale yetiştiriciliği konusunda dünyanın halen en büyük yetiştiricisi ve pazarıdır.

Asya'dan Avrupa'ya giden lale birkaç yıldır İstanbul Belediyesinin çalışmaları ile tekrar Türkiye'de boy gösteriyor. Özellikle Nisan ayı geldiğinde istanbul'un cadde ve parklarında rengarenk laleler gösterişli duruşlarıyla İstanbul'a güzellik katıyor. 400 yılın ardından lalelerin İstanbul'a çok ayrı bir hava kattığı kesin. Bu renk cümbüşünü yaşamanız dileğiyle.

istanbulmagazin.com

1) Bu yazıda geçen "talih yıldızı" ifadesi neyi tanımlamaktadır?

A) Laleye verilen isimlerden birini
B) Sevinç hissi veren laleleri
C) Şansına güvenen kişileri
D) Lale çeşitlerinden birini

2) 5. paragrafta yazar, "lale çılgınlığının en üst boyutlara ulaşması" derken neyi kastetmektedir?

A) Bazı lale türlerinin polis tarafından korunduğunu
B) Lale türlerine olan ilginin çok yüksek olmasını
C) Çok fazla insanın lale yetiştiriciliği yapmasını
D) Lale meraklılarının gizlice bahçelerinin aranmasını

3) 5. paragrafta yazar neden bir lale hırsızlığı olayından bahsetmiştir?

A) Bazı lale türlerinin çok az bulunur ve değerli olduğuna örnek vermek için
B) Polislerin o dönemde nasıl çalıştığını anlatmak için
C) Hırsızlığın o dönemde çok yaygın olduğuna vurgu yapmak için
D) Damat İbrahim Paşa'nın şüphelileri ustalıkla aradığına delil sunmak için

4) 6. paragrafa göre, aşağıdakilerden hangisi doğrudur?

A) Lale soğanları ilk olarak Hollanda'da yetiştirilmiştir.
B) Hollanda'da lale endüstrisi Leiden Üniversitesi öncülüğünde başlamıştır.
C) Kanuni Sultan Süleyman, Leiden Üniversitesi'ne lale göndermiştir.
D) Lale, Asya'dan sonra Avrupa'da da çok değerli hale gelmiştir.

5) 6. paragrafta, "lale soğanları gönderdiği sanılan kişi" kimdir?

A) Avusturya-Macaristan İmparatoru
B) Kanuni Sultan Süleyman
C) Ogier Ghislain de Busbeck
D) Carolus Clusius

6) Bu yazıda, yazarın amacı genel olarak aşağıdakilerden hangisidir?

A) Lalenin İstanbul'u nasıl güzelleştirdiğini anlatmak
B) Lalenin tarihi hakkında kısa bilgi vermek
C) Değerli lale türlerinin nasıl üretildiğini açıklamak
D) Osmanlı halkının laleye olan düşkünlüğünü vurgulamak

7) "Lale konusunda çılgınlık yapanlar hiç de az değildi."
Bu cümlenin aşağıdaki paragrafta A, B, C, D noktalarından hangisine yerleştirilmesi uygundur?

(A) Bu çılgınların arasında saray çalışanlarından çiftçisine kadar birçok kişi vardı. (**B)** Osmanlı vezirleri, sadrazamları lalelerle ilmî olarak ilgileniyor, nadide bir lâle soğanına servet ödendiğini bilenler ise, yeni bir çeşit bulmanın hayaliyle bahçelerinde gizli gizli deneyler yapıyordu. **(C)** Ardı ardına yeni lale çeşitleri çıkmış, bunlara da görüntülerinin güzelliğine yakışır, pırıltılı adlar takılmıştı. **(D)** Kimileri laleye "gönül yakan" adını vermeyi uygun görmüş, kimisi şans getireceğine inanarak "talih yıldızı" demiş, bazıları da hissettiklerine tercüman olması için, "sevinç ışığı" adını vermişti çiçeğine.

A) A B) B
C) C D) D

8) Bu parçanın özeti için giriş cümlesi aşağıda verilmiştir. Özetin tamamlanması için verilen 6 cümle arasından en uygun 3 cümle seçiniz.

• Son zamanlarda İstanbul'un parklarını süsleyen lalenin Osmanlı'ya kadar dayanan bir geçmişi vardır.
•
•
•

1) Bazı insanlar az bulunan lale türlerini bahçelerinde gizlice yetiştirmişlerdir.

2) Çeşitli renklerde yaprakları olan nadir lale türleri İstanbul'a da renk katmıştır.

3)Lalenin Avrupa'da da yayılması çok zaman almamış ve günümüzde Hollanda dünyanın en büyük lale yetiştiricisi haline gelmiştir.

4) Osmanlı insanlarının da ilgisiyle yeni lale türleri bulunmuş ve nadide lale türleri çok değerli hale gelmiştir.

5) Taç-ı Kayser, Çırağan Sarayı'nın gelmiş geçmiş en değerli lalesi olmuştur.

6) İstanbul'un fethinin ardından lale giderek popüler olmuş ve birçok kişinin uğraşı haline gelmiştir.

A) 1-3-4
B) 3-4-6
C) 5-1-6
D) 2-6-5

Bölüm 4 (8 Soru)

Aşağıdaki metni okuyarak her soru için doğru cevabı işaretleyiniz.

Beyoğlu'nun En Güzel Abisi

Taksi, yeniden başlayan kar yağışının altında yer yer buz tutmuş asfaltta ağır ağır ilerleyerek Tepebaşı'ndan aşağı iniyordu. Benim emektarı erken bir saatte emniyetin bahçesinde bırakmıştım, hayır, bu defa arıza yapan o değil, bendim. Başım dönüyordu biraz; yorgunluk mu, gerginlik mi, yoksa bedenimi sinsice ele geçirmeye başlayan yaşlılık mı, arada bir böyle oluyordu işte.

Gözlerim yarı kapalı dışarıdaki karanlığı izlerken aklım hala Kudret'in avukatı Sacit'le meşguldü. Bir zamanların gözde hukukçularından Sacit Kasımoğlu'yla. Bu adamı anlamak için geçmişini iyi bilmek gerekiyordu. İstanbul Hukuk Fakültesi'nde okumuş, yüksek lisansını Sorbonne'da yapmış olan nam-ı diğer Damat Sacit'le. Oldukça eskilerden tanıyordum onu. Sadece avukatlığının şaşaalı döneminden değil, politik olarak da en kudretli olduğu günlerden. Damat lakabı, dönemin bakanlarından birinin kızıyla evlenmesinden geliyordu. Günahı söyleyenin boynuna, 80'li yılların başında Tarlabaşı Bulvarı açılırken vurmuş en büyük voliyi. İstanbul'un taşı toprağı altın, biz nasiplenmezsek başkaları nasiplenir diyen soysuzlar var ya işte, onların en maharetlilerinden biriydi. Ankara'yla sıkı bağı olan iş bilir takımından, suçu kitabına uydurmayı marifet sayan hukuk cambazlarından. Ama ilahi adalet mi desek, şehrin ahı mı, karısı bir kokain partisinde uygunsuz vaziyette yakalanınca şansı ters dönmüştü. Hayır, tabii ki boşamamıştı karısını. Bunlar kuru iftira diyerek haberi yapan gazeteci kadını, önce dava etmiş tutturamayınca da ayağından vurdurtmuştu. Fakat gazeteci sağlam çıkmış, peşini bırakmamıştı; meslektaşının da yardımıyla kirli dosyalar tek tek açılmaya başlanmıştı.

Felaketler geldi mi peş peşe gelir derler ya, aynen öyle olmuş, kayınpederinin partisi seçimlerde iktidarı kaybetmişti. Yeni oluşan mecliste yolsuzlukla suçlanan kayınpeder, ince politik ayarlarla paçayı sıyırıp kapağı Kanada'ya atınca zaten pek de sağlam olmayan evliliği iyice çatırdamaya başlamış, çok geçmeden de magazin basınının değişmez kahramanı olan eşi soluğu babasının yanında almıştı. Eğer malına mülküne haciz gelmeseydi tanınmış avukatımız hiç itiraz etmeyecekti bu duruma. Ama yeni hükümetin şimşeklerini üzerine çekmiş bulunuyordu, kayınpederin gizli ortağı olduğu düşünülen adalet savunucumuz, böylece elinde avucunda ne varsa hepsini kaybetmişti. Daha da beteri, eriyen servetiyle birlikte itibarının da yok olmasıydı. Saygın ve elbette paralı müşterileri birer birer bıraktılar Sacit'i. Böylece o parlak avukat Kara Nizam gibi orta boy mafya babalarının savunuculuğuna kadar düştü. Ama ne yalan söyleyeyim, hiçbir zaman kibarlığından ödün vermedi. Kötülüğü de, rezilliği de hep belli bir zarafet içinde yapmayı sürdürdü. Hiçbir zaman yenilmiş biri gibi davranmadı. Bu gece de sorgu odasına aynı özgüvenle girmişti. Oysa hatalıydı, geç kalmış, müvekkilini zor durumda bırakmıştı. Hiç umurunda değilmiş gibi gülümseyerek selamlamıştı hepimizi. Vitrini de yerindeydi doğrusu, yeni yaptırdığı dişleri ağzına biraz büyük gelse de özenle taranmış sonradan ekilme saçları, kaliteli kumaştan siyah paltosu, lacivert takım elbisesi, vişneçürüğü rengindeki kravatı ve elindeki halis deriden çantasıyla zımba gibi bir adalet savaşçısı olarak dikilmişti karşımıza. Herkese iyi geceler diledikten sonra paltosunu katlayarak oturacağı iskemlenin arkasına koymuştu. Sanki ilk kez fark ediyormuş gibi, ela gözlerini sahte bir şaşkınlıkla iri iri açıp müvekkiline bakmıştı.

Ahmet Ümit

1) Yazarın taksi kullanmasının sebebi .....

A) emniyete geç kalmak istememesidir
B) yaşlanmış olmasıdır
C) yazarın kendini iyi hissetmemesidir
D) karda yürüyememesidir

2) Bu yazıda geçen "benim emektarı" ifadesi neyi kastetmektedir?

A) yazarın kendisini
B) yazarın emekli babasını
C) yazarın bindiği taksiyi
D) yazarın arabasını

3) Bu yazıdan anlaşıldığı üzere, Sacit ..... .

A) eskiden başarılı bir avukattı
B) politikacılarla iyi geçinemezdi
C) hukuk fakültesinde yüksek lisans yapmıştı
D) bir dönem bakanlık yapmıştı

4) Sacit'in Tarlabaşı Bulvarı açılırken büyük voli vurması .....

A) 1980 yılında olmuştur
B) bir bakan sayesinde olmuştur
C) kesin bir bilgi değildir
D) maharetli hukukçular sayesinde

5) Yazara göre, Sacit'in karısını boşamamasının sebebi ..... .

A) karısına iftira atılmasıdır
B) politik bağlantılarını kaybetmek istememesidir
C) bir gazetecinin Sacit'e dava açmasıdır
D) kayınpederinin yakında iktidara geleceğidir

6) Sacit'in bütün mal varlığını ve itibarını kaybetmesinin asıl sebebi aşağıdakilerden hangisidir?

A) Karısının magazin haberlerinde yer alması
B) Paralı müşterilerinin Sacit'i bırakması
C) Yeni hükümetin eski yolsuzlukların üzerine gitmesi
D) Kendisi hakkında gazetelerde yalan haberler yapılması

7) Geçmiş yaşantısını onaylamamasına rağmen yazar, Sacit'in..... .

A) kibarlığı ve özgüveninden övgüyle bahsetmektedir
B) hatasını kabul etmesini saygıyla karşılamaktadır
C) onca olumsuzluğu olgunlukla karşılamasını takdir etmektedir
D) dış görünüşüne gösterdiği özene hep imrenmektedir

8) Yazar bu hikayeyi okuyucuya niçin anlatmaktadır?

A) İyi bir hukukçuyla kötü bir hukukçu farkına örnek göstermek
B) Sacit'in, arkadaşı Kudret'e uygun bir avukat olmadığını göstermek
C) Romanda bir hukukçunun hayatını ele alacağını okuyucuya iletmek
D) Romandaki bir karakterin geçmişi hakkında bilgi vermek

Bölüm 5 (7 Soru)

Aşağıdaki metni okuyunuz ve verilen şeklin kısımlarını üzerlerine yazınız.

Yaprağın Yapısı

Yapraklar genellikle gövde üzerinde bulunan düğüm adı verilen şişkin bölgelerden çıkan, fotosentez işlevini yerine getiren geniş, yassı ve yayvan yapılardır. Yaprakların damar şekilleri, damarların paralel veya ağsı, yaprağın etli, bütün veya parçalı oluşu, saplı veya sapsız oluşu, saplarında stipullarının bulunup bulunmaması, kenarlarının düz veya dişli oluşu, kışın dökülüp dökülmediği, tüylü veya tüysüz oluşu, bitkilerin teşhis ve sınıflandırmalarında önemli rol oynar.

Bir yaprağın bazı kısımları

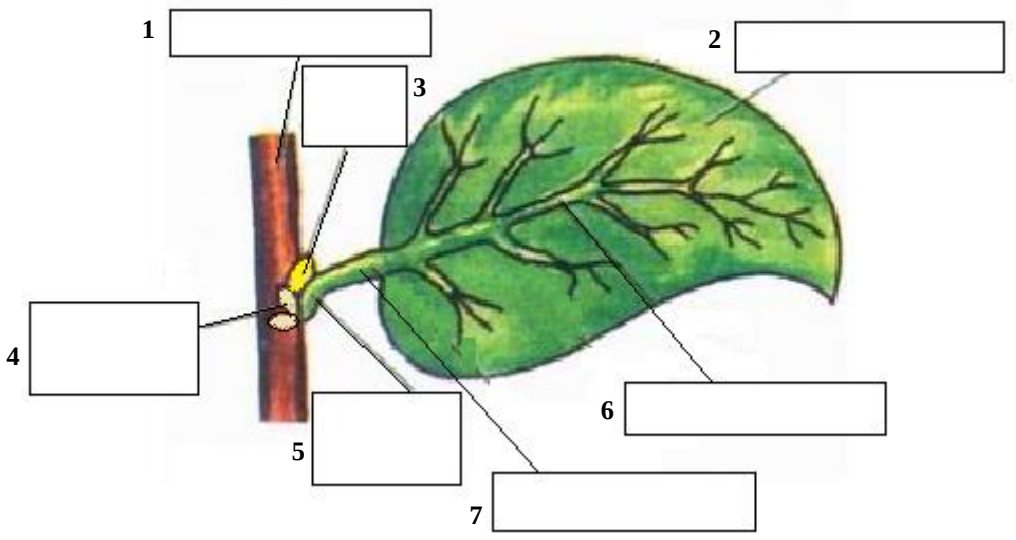
Yaprak ayası: Esas yaprağı meydana getiren geniş kısımdır. Genellikle üst yüzü yeşil, alt yüzü daha soluk yeşil renkte ve çoğunlukla iki simetrik parçadan meydana gelmiştir.

Yaprak sapı: Yaprak ayasını gövdeye, yani 'dal'a bağlayan kısımdır. Yaprak sapı çok kısadır. Yaprak sapı bulunmadığı zaman yaprak doğrudan gövdeye bağlanır. Petiolün kaidesi (noda bağlandığı yer), farklı şekillerde olabilir. Bazen genişlemiş olup 'yastıkçık ' adını alır.

Yaprak tabanı: Sapın gövde ile birleşen ve saptan daha geniş olan kısımdır. Yaprak tabanı yaprak sapı ve gövdenin birleştiği koltukta bulunan tomurcukları koruyacak şekilde genişleyerek bunları sardığı takdirde yaprak kını adını alır. Yaprak ayasında iletimi sağlayan iletim borularına ise "damarlar" denir.

Ayrıca yaprağın dala bağlandığı yerde, iki adet küçük yaprakçık bulunur. Bu yapıya kulakçık denir.

turkiyebitkileri.com



APPENDIX C

ITEM SPECIFICATIONS

General concerns

Addressees: Individuals studying in a university in Turkey and learning Turkish as a foreign language.

Operations: Expeditious reading: Skim for main ideas; search read for information; scan for specific items.
Careful reading: Identify main ideas, author purpose, identify implicit information; guess the meaning of unfamiliar words from the context; distinguish between fact and opinion.

Reading Levels: The word ‘local’ in the following item specifications means that the indicated items are intended to require reading one sentence or a few sentences at most. The word ‘global’ means that the indicated items are intended to require reading one paragraph, multiple paragraphs or the whole text.

Item Specifications for Task 1 (B1)

Strand: Reading
Sub-strand: Comprehension
Standard: Students will read an expository text, using careful and expeditious reading strategies and will demonstrate literal and interpretive comprehension.

Contextual Features
Length: 300-350 words.
Vocabulary: Average characters per word: 5-7. Prefer texts with mostly frequent vocabulary. Very infrequent words can be provided with definitions.
Grammar: Average words per sentence: 9-12. Complex sentences and embedded clauses can appear only rarely.
Rhetorical organization: Prefer texts with explicit flow of information and clear organization.
Genre: Magazine or newspaper article. The text should be free from journalistic style and organization. The text should be edited to adjust style, organization and clarity.
Subject specificity: Prefer texts that do not require subject specific knowledge.
Cultural specificity: Prefer texts that do not require cultural knowledge.
Text abstractness: Prefer concrete subjects. Majority of the content words should be concrete.
Number of Items: 6-8.
Items below represent possible tasks that can be used in a test with similar purposes.

Item Benchmarks	Intended Reading Type/Level	Item Type
Item 1 Test takers will carefully read one/two sentence(s) and identify explicit details from the text	Careful - Local	Multiple choice
Item 2 Test takers will scan part of the text and identify explicit details	Expeditious - Local	Multiple choice
Item 3 Test takers will carefully read one/two sentence(s) and identify explicit details from the text	Careful - Local	Multiple choice
Item 4 Test takers will carefully read one/two sentence(s) and identify explicit details from the text	Careful - Local	Multiple choice
Item 5 Test takers will carefully read one/two sentence(s) and identify explicit details from the text	Careful - Local	Multiple choice
Item 6 Test takers will identify the author's purpose based on the explicit and implicit information from the text	Expeditious - Global	Multiple choice

Item Specifications for Task 2 (B2)

Strand: Reading

Sub strand: Comprehension

Standard: Students will read an expository text, using careful reading strategy and will demonstrate literal, interpretive, inferential and evaluative comprehension.

Contextual Features

Length: 300-350 words.

Vocabulary: Average characters per word: 6-8. Prefer texts with mostly frequent vocabulary. Infrequent words can appear occasionally.

Grammar: Average words per sentence: 11-14. Complex sentences and embedded clauses can appear occasionally.

Rhetorical organization: Prefer texts with mainly explicit flow of information and clear organization.

Genre: Magazine or newspaper article. The text should be free from journalistic style and organization. The text should be edited to adjust style, organization and clarity.

Subject specificity: Prefer texts that do not require subject specific knowledge.

Cultural specificity: Prefer texts that do not require cultural knowledge.

Text abstractness: Prefer concrete subjects. Majority of the content words should be concrete.

Number of Items: 8-10.

Items below represent possible tasks that can be used in a test with similar purposes.

Item Benchmarks	Intended Reading Type/Level	Item Type
Item 1 Test takers will distinguish between fact, opinion and non-existent information in the text	Careful - Local	Multiple choice
Item 2 Test takers will distinguish between fact, opinion and non-existent information in the text	Careful - Local	Multiple choice
Item 3 Test takers will distinguish between fact, opinion and non-existent information in the text	Careful - Local	Multiple choice
Item 4 Test takers will distinguish between fact, opinion and non-existent information in the text	Careful - Local	Multiple choice
Item 5 Test takers will distinguish between fact, opinion and non-existent information in the text	Careful - Local	Multiple choice
Item 6 Test takers will distinguish between fact, opinion and non-existent information in the text	Careful - Local	Multiple choice
Item 7 Test takers will distinguish between fact, opinion and non-existent information in the text	Careful - Local	Multiple choice
Item 8 Test takers will distinguish between fact, opinion and non-existent information in the text	Careful - Local	Multiple choice
Item 9 Test takers will distinguish between fact, opinion and non-existent information in the text	Careful - Local	Multiple choice
Item 10 Test takers will distinguish between fact, opinion and non-existent information in the text	Careful - Local	Multiple choice

Item Specifications for Task 3 (C1)

Strand: Reading

Sub strand: Comprehension

Standard: Students will read an expository text, using careful and expeditious reading strategies and will demonstrate literal, interpretive, inferential and evaluative comprehension.

Contextual Features

Length: 400-450 words.

Vocabulary: Average characters per word: 6-8. Infrequent words, metaphors and idiomatic expressions can often appear in the text.

Grammar: Average words per sentence: 12-15. Complex sentences and embedded clauses can appear frequently.

Rhetorical organization: The flow of information in the text can be both explicit and implicit. Organization can sometimes be blurry. Time and order of events in the text can occasionally change back and forth.

Genre: Magazine or newspaper article. The text can have journalistic style and organization. Minor adjustments can be done on style, organization and clarity.

Subject specificity: The text can contain subject specific information; however it should be comprehensible by means of the explanations provided in the text.

Cultural specificity: The text can contain cultural elements; however they should be comprehensible by means of the information provided in the text.

Text abstractness: The subject can be an abstract one. Abstract content words can appear frequently in the text.

Number of Items: 8-10.

Items below represent possible tasks that can be used in a test with similar purposes.

Item Benchmarks	Intended Reading Type/Level	Item Type
Item 1 Test takers will make inferences and draw accurate conclusions based on explicit information from the text	Careful - Local	Multiple choice
Item 2 Test takers will determine the meaning of idiomatic expressions from the context	Careful - Local	Multiple choice
Item 3 Test takers will make inferences and draw conclusions based on explicit information from the text	Careful - Global	Multiple choice
Item 4 Test takers will read the text carefully and identify implicit details from the text	Careful - Global	Multiple choice
Item 5 Test takers will scan the text and identify explicit details from the text	Expeditious - Local	Multiple choice
Item 6 Test takers will identify the author's purpose based on the explicit and implicit information from the text	Careful - Global	Multiple choice
Item 7 Test takers will read the text carefully and process the cohesion of the text	Careful - Global	Multiple choice
Item 8 Test takers will recognize the order of main ideas in the text	Expeditious - Global Careful - Global	Sentence ordering

Item Specifications for Task 4 (C2)

Strand: Reading

Sub strand: Comprehension

Standard: Students will read a literary text, using careful reading strategy and will demonstrate literal, interpretive, inferential and evaluative comprehension.

Contextual Features

Length: 400-450 words.

Vocabulary: Average characters per word: 6-9. Infrequent words can appear very often in the text. The text may contain metaphors and idiomatic expressions with nearly no restriction. Too technical or specialized words can appear with explanations.

Grammar: Average words per sentence: 13-16. Complex sentences and embedded clauses can appear all through the text.

Rhetorical organization: The flow of information in the text can be both explicit and implicit. Organization does not have to be clear. Change in the time and order of events in the text is quite flexible.

Genre: Excerpt from a novel or short story

Subject specificity: The text can contain subject specific information; however it should be comprehensible by means of the explanations provided in the text.

Cultural specificity: The text can contain cultural elements; however they should be comprehensible by means of the information provided in the text.

Text abstractness: The subject can be an abstract one. Abstract content words can appear all the time with no restriction.

Number of Items: 8-10.

Items below represent possible tasks that can be used in a test with similar purposes.

Item Benchmarks	Intended Reading Type/Level	Item Type
Item 1 Test takers will make inferences and draw accurate conclusions based on explicit information from the text	Careful - Local	Multiple choice
Item 2 Test takers will carefully read one/two sentence(s) and identify explicit details from the text	Careful - Local	Multiple choice
Item 3 Test takers will carefully read the text and identify explicit and implicit details from the text	Careful - Global	Multiple choice
Item 4 Test takers will carefully read the text and identify explicit and implicit details from the text	Careful - Local	Multiple choice
Item 5 Test takers will carefully read the text and identify implicit details from the text	Careful - Global	Multiple choice
Item 6 Test takers will make inferences and draw conclusions based on explicit and implicit information from the text	Careful - Global	Multiple choice
Item 7 Test takers will make inferences and draw conclusions based on explicit and implicit information from the text	Careful - Global	Multiple choice
Item 8 Test takers will identify the author's purpose based on the explicit and implicit information from the text	Careful - Global	Multiple choice

APPENDIX D

READING TEST EXPERT EVALUATION FORM

Task: ….

Part A

For each of the items below, circle the number that reflects your opinion on a four-point scale where: 1 = Strongly disagree, 2 = Disagree, 3 = Agree 4 = Strongly agree

INSTRUCTIONS				
Instructions are clear.	1	2	3	4
Instructions are adequate.	1	2	3	4
Instructions are relevant.	1	2	3	4
QUESTIONS				
The questions are clear.	1	2	3	4
The language of items is easier than the language of the text.	1	2	3	4
The questions can only be answered if the text is read.	1	2	3	4
TEXT				
The text is appropriate in an academic context.	1	2	3	4
The text length is appropriate in an academic context.	1	2	3	4
The text does not require high levels of knowledge to comprehend.	1	2	3	4
The text does not require cultural knowledge to comprehend.	1	2	3	4

Part B

In your view, which of the following skill(s) does each item measure?

	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10
Skimming for overall gist										
Demonstrating understanding of text as a whole										
Identifying topic of text										
Identifying function of text										
Distinguishing main points of text from subsidiary ones										
Retrieving specific information by scanning text										
Locating and selecting relevant factual information to perform task										
Demonstrating understanding of how text structure works										
Distinguishing fact from opinion										
Deducing meaning from context										
Interpreting text for author's attitude, style										
Making inferences from information given in the text										
Making use of clues such as subtitles, illustrations										
Other (please specify)										

Part C

If you have any suggestions to improve any of the questions, please indicate them below.

_____
_____
_____
_____
_____
_____
_____
_____

APPENDIX E

READING TASKS AFTER REVISION

OKUMA - ANLAMA

Bölüm 1 (6 Soru) **Süre: 15 dakika**

Aşağıdaki metni dikkatlice okuyarak her soru için doğru yanıtı işaretleyiniz.

Köpekbalıklarının Soyu Tükeniyor

1 Yeni yapılan bir araştırmaya göre okyanuslardaki köpekbalıklarının yarısından çoğu, soylarının tükenmesi tehlikesiyle karşı karşıya. Çeşitli bilimsel gruplar bu tehlikenin nedenlerini ve çözümlerini tartışıyorlar. Uluslararası Doğa Koruma Birliği'nden (IUCN) uzmanlar, 11 köpekbalığı türünün yüksek risk listesinde olduğunu söylüyorlar. Üstelik 5 türün daha bu listeye girme olasılığı var. Köpekbalıkları çok yavaş ürüyor ve bu nedenle aşırı avlanmadan çok etkileniyorlar. Bilim insanları köpekbalıkları için küresel avlanma sınırlamaları getirilmesini istiyorlar. Buna ek olarak, yüzgeçleri[1] için avlanmalarına son verilmesini ve hata sonucu yakalanmalarının önlenmesini istiyorlar.

2 IUCN Köpekbalığı Uzman Grubu'ndan Sonja Fordham, "İnsanlar köpekbalıklarının aşırı avlanmaya karşı dirençli olduğunu sanıyor ama bu doğru değil" diyor. Fordham ayrıca şunları da ekliyor "Köpekbalıkları için uluslararası yakalama sınırları getirilmiyor. Bu nedenle de giderek risk altına giriyorlar. Okyanuslarda balıkçılığın yoğun yapıldığı alanlar var ve oralardaki köpekbalıkları çoğunlukla korunmasız."

Yeni Tehditler

3 Bilim insanları, 21 tür köpekbalığına yönelik araştırmaların verilerini değerlendirdi. Bu değerlendirmeye göre, 21 tür içinden biri, dev şeytan vatozları, soyu tükenmek üzere, 10 türde de büyük bir azalma var. Geri kalan beşindeyse azalma oranı çok ciddi değil. Bu türler 'tükenmeye yakın' olarak tanımlanıyor.

Yüzgeç Kesimi

4 Köpekbalıklarına yönelik en önemli tehdit, bilerek ya da yanlışlıkla yapılan avlanma. Fordham, "Köpekbalıkları, önceleri kılıçbalıklarını avlayan gemilerce yanlışlıkla yakalanıyordu. Ama şimdi köpekbalıkları balıkçıların özellikle hedefi oluyor. Bazı türler yüzgeçleri ve eti için, bazı türler de yalnızca yüzgeçleri için avlanıyor." diyor.

5 Uluslararası sularda köpekbalığı avcılığını düzenleyen birçok organizasyon var. Bu organizasyonlar, köpekbalıklarının sadece yüzgeçleri için avlanmalarını sınırlamak istiyorama bu organizasyonlardan her biri değişik standartlar uyguluyor. Bu da avcıların yasaları çiğnemelerine olanak tanıyor. Koruma grupları, Doğu Asya ülkelerinin ekonomileri büyüdükçe, köpekbalıklarının yüzgeçleri için daha çok avlandığınısöylüyor. IUCN raporunun baş yazarı, Simon Fraser Üniversitesi'nden Nicholas Dulvy "Balıkçılık yetkililerine ve konuyla ilgili bölgesel, ulusal ve uluslararası yetkililere bu durumu düzeltmek için büyük görev düşüyor. Bu gruplar politik olarak kararlı davranmalıdır. Ama herşeyin ötesinde güçlü bir halk desteği gerekir. Yerel halk desteklerse bu sorun kolaylıkla çözülebilir. "

6 Rapor, Bonn'da yapılan Biyolojik Çeşitlilik Konvensiyonu'nda sunuldu ve Deniz ve Tatlı Su Ekosistemleri adlı dergide bu yılın sonunda yayımlanacak. Dergide yeni risk değerlendirmeleri de yayımlanacak.

Korkut Demirbaş
Bilim ve Teknik 4 Temmuz 2008

Yüzgeç: Balığın su içerisindeki hareketine yardımcı olan organlardır. Özellikle Asya ülkelerinde bazı köpek balığı türleri yüzgeçleri için avlanmaktadır. Bunun en büyük nedeni ise köpek balığı yüzgecinden yapılan çorbanın oldukça pahalı olmasıdır.

1) 1. paragrafta bilim insanlarının, soyları tehlikede olan köpekbalıklarını kurtarmak için önerdiği bazı düzenlemeler vardır. Aşağıdakilerden hangisi bu düzenlemelerden biri DEĞİLDİR?

A) Uluslararası avlanmanın kontrol edilmesi
B) Yüzgeçleri için avlanmasının yasaklanması
C) Yanlışlıkla yakalamanın azaltılması
D) Daha fazla üremeleri için uygun şartların yaratılması

2) Metne göre kaç köpek balığı türü yok olma riski altındadır?

A) 11
B) 5
C) 21
D) 10

3) 4. paragrafa göre,eskiden balıkçıların köpekbalığı yakalamasıyla ilgili aşağıdakilerden hangisi doğrudur?

A) Yanlışlıkla yapılırdı.
B) Onları çok uğraştırırdı.
C) Sadece etleri için yapılırdı.
D) Diğer balıklar azaldığında olurdu.

4) 5. paragrafa göre köpekbalığı avcılığını sınırlama çabaları neden işe yaramıyor?

A) Farklı uygulamalar olduğu için.
B) Köpekbalığı yüzgeçleri popüler olduğu için.
C) Bazı köpekbalığı türleri risk altında olmadığı için.
D) Hükümetler bu konuyu ciddiye almadığı için.

5) Nicholas Dulvy'e göre köpekbalığı türlerini yok olma tehlikesinden kurtarmak için gereken en önemli şey nedir?

A) Köpekbalığına dayalı ekonomilerin küçültülmesi
B) Balıkçılara eğitimler verilmesi
C) Güçlü bir halk desteği
D) Yetkililerin politik görüşlerini değiştirmesi

6) Bu yazıda yazarın genel amacı aşağıdakilerden hangisidir?

A) Köpekbalıklarını olumsuz etkileyen çevresel faktörleri tartışmak ve çözümler sunmak
B) Köpek balıklarının sayılarının neden azaldıkları ve bu konuda yapılabilecekler konusunda bilgi vermek
C) Biyolojik çeşitlilik konusunda yapılan bilimsel araştırmaları açıklamak ve yeni araştırmaları cesaretlendirmek
D) Modern ekonomilerin köpek balıklarına büyük zarar verdiğini tartışmak ve

Bölüm 2 (10 Soru) **Süre: 15 dakika**

Aşağıdaki metni okuyarak her soru için doğru yanıtı işaretleyiniz.

Hoffman'a Veda

1 Oynadığı her filme kendi duygularını katan, filmde kendi varlığını hissettiren oyunculardan olan Philip Seymour Hoffman artık aramızda değil. 2 Şubat günü evinde, kolunda bir şırıngayla ölü bulundu. Malum, aşırı doz. Bu alçak gönüllü ve mağdur oyuncunun bu şekilde, henüz kırk altı yaşındayken ölmesi üzücü fakat şaşırtıcı değil. Canlandırdığı hafiften kaybeden, hayatın kenarında duran, trajik ve krizli karakterlere uygun bir son. Yani bir bakıma Hoffman, yaşadığı gibi oynamış, oynadığı gibi de hayata veda etmiş oldu.

2 Hoffman gerçek bir 'yıldız'dı ama 'yıldız' sistemine uygun bir tavra ve duruşa sahip değildi. Aslında 'silik' olabilecek olağan hatları, hafif tombulluğu, hantallığı ve döküklüğüyle sorunlu 'yan komşu' karakterlerine ve daha doğrusu genel olarak yan rollere uygundu. Ama yan rollerde belirdiğinde bile insanın aklına bir çengel gibi takılıyor, başrolden 'rol' çalıyordu. Hoffman etkileyici bir fiziğe sahip değildi. Karizması, çok insani ve çok sıradan ama bir o kadar da sahici bir portre çizmesinden geliyordu. İnsanda içten samimi bir sevgi yaratmasının nedeni de buydu herhalde.

3 Bu hakikilik sayesinde 'iyi insan' rollerini de 'kötü insan' rollerini de müthiş bir sahicilikle canlandırıyordu. Aslında tıpkı yan rol/baş rol ayrımını ortadan kaldırdığı gibi iyi/kötü ayrımını da ortadan kaldırıyordu. Belki de bu yüzden, bir kategoriye kolayca yerleştirilip kenara koyulamadığı için filmlerdeki varlığı seyircilerin hep aklında kalıyordu.

4 Hoffman, Paul Thomas Anderson'ın neredeyse bütün filmlerinde rol aldı. Oynadığı son Anderson filminin, sonunu hazırlayan bir olaya vesile olması son derece ironik. Corinne Van Vliet'in yazısına göre, Hoffman en son 1989'da veda ettiği madde bağımlılığına, *The Master*'ın 2012'de yapılan galasında kendisine ikram edilen bir kadeh içkiyle geri dönmüş. Daha doğrusu şöyle: tam yirmi üç yıldır alkolden ve uyuşturuculardan uzak duran Hoffman ilk kez bu galada alkolü tekrar bünyesine sokmuş. İşte o alkol damlası Hoffman'ı tekrar bağımlılığa götüren yolun başlangıcı olmuş. Hızla yenilenen eski bağımlılık Hoffman'ı iki seneden kısa bir süre içinde aşırı dozdan ölüme götürmüş. Bu son filmin adının 'Usta' olması da Hoffman'a uygun bir unvan olması sebebiyle hoş (ve acı) bir ironi içeriyor. Trajik ve erken ölümü de filmlerle geçen hayatına bir film sahnesi gibi eklenmiş oldu. Huzur içinde yatsın.

Ahmet Ergenç
Altyazı Dergisi

A. Aşağıdaki cümleleri okuyunuz ve her cümle için F, G ve Y'den birini işaretleyiniz.

F: Cümle yazarın kişisel fikriyse
G: Cümle nesnel bir gerçekten bahsediyorsa
Y: Cümle bu yazıda yer almayan bir ifadeyse

1. Hoffman'ın ölüm sebebi uyuşturucudur.	**F**	**G**	**Y**
2. Hoffman'ın ölümü canlandırdığı karakterlere uygun bir son oldu.	**F**	**G**	**Y**
3. Hoffman neşeli, hayat dolu karakterleri canlandırmak istemiştir.	**F**	**G**	**Y**
4. Hoffman oynadığı rollere samimiyet katabiliyor, bu da onu etkileyici bir oyuncu yapıyordu.	**F**	**G**	**Y**
5. Oynadığı rollerden dolayı Hoffman'ı hemen etiketlemek zor olduğu için seyircinin aklında kalıyordu.	**F**	**G**	**Y**
6. Hoffman yan rollere kattığı hakikilikle bir yıldızdı.	**F**	**G**	**Y**
7. Hoffman, yönetmen Anderson'un filmlerinin çoğunda rol alıyordu.	**F**	**G**	**Y**
8. Hoffman bağımlılıktan kurtulmak için 2012 yılına kadar tedavi görmüştür.	**F**	**G**	**Y**
9. "The Master" filminin galasına kadar Hoffman, çok uzun bir süredir alkol kullanmıyordu.	**F**	**G**	**Y**
10. Hoffman'ın bağımlılığı ailevi sebeplerden kaynaklanıyordu.	**F**	**G**	**Y**

Bölüm 3 (8 Soru) **Süre: 20 dakika**

Aşağıdaki metni okuyarak her soru için doğru yanıtı işaretleyiniz.

Lale

1 Lale zambakgiller ailesinden, yaprakları uzun ve mızraksı, çiçekleri kadeh biçiminde, türlü renkte, alacalı bir süs bitkisidir. Laleler şimdilerde, birbirinden farklı renkleri, zerafetleriyle İstanbul'un cadde ve parklarını süslüyorlar. Özellikle son zamanlarda tekrar tanıştığımız lale, hiçbir çiçekte olmadığı kadar insanları etkisi altına almış ve renkli bir öykünün baş kahramanı olmuştu.

2 İnsanları hastalık derecesinde kendisine tutkun eden lalelerin öyküsü Osmanlılarla başlamıştır. İstanbul'un Fethi'nden sonra, Fatih'in emri ile yeniden düzenlenen bahçeler (parklar) lâlelerle süslenmiştir. Zaten Fatih Sultan Mehmet bir bahçıvandı. Boş vakitlerinin çoğunu bunun için harcar ve bundan büyük bir haz duyardı. Boş zamanlarda Topkapı ve diğer sarayların bahçelerinde çalışmaktan da büyük zevk alırdı.

3 Kanuni devrinde de, lale türleri geliştirip çoğaltılmıştır. 16. yüzyılda İstanbul'da yayılıp, oradan da Avrupa'ya kadar sıçramıştı. Muhteşem Süleyman döneminde küçük ve kısa boylu, çiçeği badem biçiminde, uçları ince ve sivri, İstanbul'a özgü lâle çeşitleri yetiştirilmiş, bilmeden de olsa lale çılgınlığı böylece başlatılmıştı. İnce mi ince, uzun mu uzun laleler öylesine nazlı görünüyorlardı ki, bir kere gören etkisinden kurtulamıyor, böylesi bir güzelliğe sahip olabilmek için akla gelmez çılgınlıklar yapılıyordu.

4 Kimler yoktu ki, bu çılgınların arasında. Bazıları lalelerle ilmî olarak ilgileniyor, nadide bir lâle soğanına servet ödendiğini bilenler ise, yeni bir çeşit bulmanın hayaliyle bahçelerinde gizli gizli deneyler yapıyordu. Ardı ardına yeni lale çeşitleri çıkmış, bunlara da görüntülerinin güzelliğine yakışır, pırıltılı adlar takılmıştı. Kimileri laleye "gönül yakan" adını vermeyi uygun görmüş, kimisi şans getireceğine inanarak "**talih yıldızı**" demiş, bazıları da hissettiklerine tercüman olması için, "sevinç ışığı" adını vermişti çiçeğine.

5 Ancak laleler sadece güzelliklerle anılmıyor, kimi zaman da polisiye olaylarla gündeme geliyordu. Örneğin lale çılgınlığının en üst boyutlara ulaştığı Lale Devri'nde, Taç-ı Kayser adı verileren nadir bir lale türü Çırağan Sarayı'nın bahçesine ekilmiş ancak çalınmıştı. Damat İbrahim Paşa, işin peşini bırakmamış. İlk şüpheliler lale tutkunları olduğu için, gizlice bahçeleri aranmış. Nafile ki tüm çabalar boşa çıkmıştı.

6 Avusturya-Macaristan İmparatorluğu'nun Kanuni Sultan Süleyman nezdindeki büyükelçisi Ogier Ghislain de Busbeck'in 1554 yılında geldiği İstanbul'dan Avusturya'da yaşayan dostu Carolus Clusius'a lale soğanları gönderdiği sanılmaktadır. Daha sonra Hollanda'ya giderek Leiden Üniversitesi'nde göreve başlayan Clusius, bu ülkelerde laleyi ilk yetiştiren ve lâle endüstrisini kuran kişi olarak bilinmektedir. Avrupa'ya giden lale, özellikle Hollanda ve Almanya'da aranan bir meta haline gelmişti ve bir lale soğanına bütün servetini yatıranlar vardı. Hollanda, günümüze kadar gelen lale yetiştiriciliği konusunda dünyanın halen en büyük yetiştiricisi ve pazarıdır.

7 Asya'dan Avrupa'ya giden lale birkaç yıldır İstanbul Belediyesinin çalışmaları ile tekrar Türkiye'de boy gösteriyor. Özellikle Nisan ayı geldiğinde İstanbul'un cadde ve parklarında rengarenk laleler gösterişli duruşlarıyla İstanbul'a güzellik katıyor. 400 yılın ardından lalelerin İstanbul'a çok ayrı bir hava kattığı kesin. Bu renk cümbüşünü yaşamanız dileğiyle.

İstanbul Magazin
istanbulmagazin.com

1) 4. paragrafta geçen “talih yıldızı” ifadesi neyi tanımlamaktadır?

A) Laleye verilen isimlerden birini
B) Sevinç hissi veren laleleri
C) Şansına güvenen kişileri
D) Şekli yıldıza benzeyen bir lale türünü

2) Metinde geçen “lale çılgınlığının en üst boyutlara ulaşması” ifadesiyle yazar neyi kastetmektedir?

A) Bazı lale türlerinin polis tarafından korunduğunu
B) Lale türlerine olan ilginin çok yüksek olmasını
C) Çok fazla insanın lale yetiştiriciliği yapmasını
D) Lale meraklılarının gizlice bahçelerinin aranmasını

3) Bu metinde yazar neden bir lale hırsızlığı olayından bahsetmiştir?

A) Bazı lale türlerinin çok az bulunur ve değerli olduğuna örnek vermek için
B) Polislerin o dönemde nasıl çalıştığını anlatmak için
C) Hırsızlığın o dönemde çok yaygın olduğuna vurgu yapmak için
D) Damat İbrahim Paşa’nın şüphelileri ustalıkla aradığına delil sunmak için

4) Metne göre aşağıdakilerden hangisi doğrudur?

A) Lale soğanları ilk olarak Hollanda’da yetiştirilmiştir.
B) Osmanlı’da lale endüstrisi üniversiteler öncülüğünde başlamıştır.
C) Kanuni Sultan Süleyman, Leiden Üniversitesi’ne lale göndermiştir.
D) Lale, Asya’dan sonra Avrupa’da da çok değerli hale gelmiştir.

5) Metne göre “lale soğanları gönderdiği sanılan kişi” kimdir?

A) Avusturya-Macaristan İmparatoru
B) Kanuni Sultan Süleyman
C) Ogier Ghislain de Busbeck
D) Carolus Clusius

6) Bu yazıda, yazarın amacı genel olarak aşağıdakilerden hangisidir?

A) Bir çiçek olarak lalenin güzelliğini vurgulamak
B) Lalenin tarihi hakkında kısa bilgi vermek
C) Lale türlerinin nasıl yayıldığını anlatmak
D) Osmanlı halkının laleye olan düşkünlüğünü vurgulamak

7) “Lale konusunda çılgınlık yapanlar çok farklı kesimdendi.”
Bu cümlenin aşağıdaki paragrafta A, B, C, D noktalarından hangisine yerleştirilmesi uygundur?

(A) Bu çılgınların arasında saray çalışanlarından çiftçisine kadar birçok kişi vardı. (**B)** Bazıları lalelerle ilmî olarak ilgileniyor, nadide bir lâle soğanına servet ödendiğini bilenler ise, yeni bir çeşit bulmanın hayaliyle bahçelerinde gizli gizli deneyler yapıyordu. **(C)** Ardı ardına yeni lale çeşitleri çıkmış, bunlara da görüntülerinin güzelliğine yakışır, pırıltılı adlar takılmıştı. **(D)** Kimileri laleye “gönül yakan” adını vermeyi uygun görmüş, kimisi şans getireceğine inanarak “talih yıldızı” demiş, bazıları da hissettiklerine tercüman olması için, “sevinç ışığı” adını vermişti çiçeğine.

A) A B) B
C) C D) D

8) Bu metnin bir özeti için 4 tane cümle karışık sırada verilmiştir. Özetin düzenlenmesi için, verilen cümlelerin doğru sırasını aşağıya yazarak gösteriniz.

I) Lalenin Avrupa’da da yayılması çok zaman almamış ve günümüzde Hollanda dünyanın en büyük lale yetiştiricisi haline gelmiştir.

II) Osmanlı insanlarının da ilgisiyle yeni lale türleri bulunmuş ve nadide lale türleri çok değerli hale gelmiştir.

III) Son zamanlarda İstanbul’un parklarını süsleyen lalenin Osmanlı’ya kadar dayanan bir geçmişi vardır.

IV) İstanbul’un fethinin ardından lale giderek popüler olmuş ve birçok kişinin uğraşı haline gelmiştir.

1) …..
2) …..
3) …..
4) …..

Bölüm 4 (8 Soru) **Süre: 20 dakika**

Aşağıdaki metni okuyarak her soru için doğru yanıtı işaretleyiniz.

Beyoğlu'nun En Güzel Abisi

Taksi, yeniden başlayan kar yağışının altında yer yer buz tutmuş asfaltta ağır ağır ilerleyerek Tepebaşı'ndan aşağı iniyordu. **Benim emektarı** erken bir saatte emniyetin bahçesinde bırakmıştım, hayır, bu defa arıza yapan o değil, bendim. Başım dönüyordu biraz; yorgunluk mu, gerginlik mi, yoksa bedenimi sinsice ele geçirmeye başlayan yaşlılık mı, arada bir böyle oluyordu işte.

1

Gözlerim yarı kapalı dışarıdaki karanlığı izlerken aklım hala Kudret'in avukatı Sacit'le meşguldü. Bir zamanların gözde hukukçularından Sacit Kasımoğlu'yla. Bu adamı anlamak için geçmişini iyi bilmek gerekiyordu. İstanbul Hukuk Fakültesi'nde okumuş, mastırını Sorbon'da yapmış olan namı diğer Damat Sacit'le. Oldukça eskilerden tanıyordum onu. Sadece avukatlığının şaşaalı döneminden değil, politik olarak da en kudretli olduğu günlerden. Damat lakabı, dönemin bakanlarından birinin kızıyla evlenmesinden geliyordu. Günahı söyleyenin boynuna, 80'li yılların başında Tarlabaşı Bulvarı açılırken vurmuş en büyük voliyi. İstanbul'un taşı toprağı altın, biz nasiplenmezsek başkaları nasiplenir diyen soysuzlar var ya işte, onların en maharetlilerinden biriydi. Ankara'yla sıkı bağı olan iş bilir takımından, suçu kitabına uydurmayı marifet sayan hukuk cambazlarından. Ama ilahi adalet mi desek, şehrin ahı mı, karısı bir kokain partisinde uygunsuz vaziyette yakalanınca şansı ters dönmüştü. Hayır, tabii ki boşamamıştı karısını. Bunlar kuru iftira diyerek haberi yapan gazeteci kadını, önce dava etmiş tutturamayınca da ayağından vurdurtmuştu. Fakat gazeteci sağlam çıkmış, peşini bırakmamıştı; meslektaşının da yardımıyla kirli dosyalar tek tek açılmaya başlanmıştı.

2

Felaketler geldi mi peş peşe gelir derler ya, aynen öyle olmuş, kayınpederinin partisi seçimlerde iktidarı kaybetmişti. Yeni oluşan mecliste yolsuzlukla suçlanan kayınpeder, ince politik ayarlarla paçayı sıyırıp kapağı Kanada'ya atınca zaten pek de sağlam olmayan evliliği iyice çatırdamaya başlamış, çok geçmeden de magazin basınının değişmez kahramanı olan eşi soluğu babasının yanında almıştı. Eğer malına mülküne haciz gelmeseydi tanınmış avukatımız hiç itiraz etmeyecekti bu duruma. Ama yeni hükümetin şimşeklerini üzerine çekmiş bulunuyordu, kayınpederin gizli ortağı olduğu düşünülen adalet savunucumuz, böylece elinde avucunda ne varsa hepsini kaybetmişti. Daha da beteri, eriyen servetiyle birlikte itibarının da yok olmasıydı. Saygın ve elbette paralı müşterileri birer birer bıraktılar Sacit'i. Böylece o parlak avukat Kara Nizam gibi orta boy mafya babalarının savunuculuğuna kadar düştü. Ama ne yalan söyleyeyim, hiçbir zaman kibarlığından ödün vermedi. Kötülüğü de, rezilliği de hep belli bir zarafet içinde yapmayı sürdürdü. Hiçbir zaman yenilmiş biri gibi davranmadı. Bu gece de sorgu odasına aynı özgüvenle girmişti. Oysa hatalıydı, geç kalmış, müvekkilini zor durumda bırakmıştı. Hiç umurunda değilmiş gibi gülümseyerek selamlamıştı hepimizi. Vitrini de yerindeydi doğrusu, yeni yaptırdığı dişleri ağzına biraz büyük gelse de özenle taranmış sonradan ekilme saçları, kaliteli kumaştan siyah paltosu, lacivert takım elbisesi, vişneçürüğü rengindeki kravatı ve elindeki halis deriden çantasıyla zımba gibi bir adalet savaşçısı olarak dikilmişti karşımıza. Herkese iyi geceler diledikten sonra paltosunu katlayarak oturacağı iskemlenin arkasına koymuştu. Sanki ilk kez fark ediyormuş gibi, ela gözlerini sahte bir şaşkınlıkla iri iri açıp müvekkiline bakmıştı.

3

Ahmet Ümit
Beyoğlu'nun En Güzel Abisi

1) Yazarın taksi kullanmasının sebebi .....

A) emniyete geç kalmak istememesidir
B) yaşlanmış olmasıdır
C) yazarın kendini iyi hissetmemesidir
D) karda yürüyememesidir

2) Bu yazıda geçen "benim emektarı" ifadesi neyi kastetmektedir?

A) yazarın kendisini
B) yazarın emekli polis arkadaşını
C) yazarın bindiği taksiyi
D) yazarın arabasını

3) Bu yazıdan anlaşıldığı üzere, Sacit ..... .

A) eskiden başarılı bir avukattı
B) politikacılarla iyi geçinemezdi
C) politika üzerine mastır yapmıştı
D) bir dönem bakanlıkta görev yapmıştı

4) Sacit'in Tarlabaşı Bulvarı açılırken büyük voli vurması ..... .

A) 1980 yılının başında olmuştur
B) bir bakan sayesinde olmuştur
C) kesin bir bilgi değildir
D) tamamen şans eseridir

5) Yazara göre, Sacit'in karısını boşamamasının sebebi ..... .

A) karısına iftira atılmasıdır
B) politik bağlantılarını kaybetmek istememesidir
C) bir gazetecinin Sacit'e dava açmasıdır
D) kayınpederinin karşı çıkmasıdır

6) Sacit'in bütün mal varlığını ve itibarını kaybetmesinin asıl sebebi aşağıdakilerden hangisidir?

A) Karısının magazin haberlerinde yer alması
B) Paralı müşterilerinin Sacit'i bırakması
C) Yeni hükümetin eski yolsuzlukların üzerine gitmesi
D) Kendisi hakkında gazetelerde yalan haberler yapılması

7) Geçmiş yaşantısını onaylamamasına rağmen yazar, Sacit'in..... .

A) kibarlığı ve özgüveninden övgüyle bahsetmektedir
B) hatasını kabul etmesini saygıyla karşılamaktadır
C) onca olumsuzluğu olgunlukla karşılamasını takdir etmektedir
D) giyimine gösterdiği özene hep imrenmektedir

8) Yazar bu hikayeyi okuyucuya niçin anlatmaktadır?

A) İyi bir hukukçuyla kötü bir hukukçu farkına örnek göstermek
B) Sacit'in, arkadaşı Kudret'e uygun bir avukat olmadığını göstermek
C) Romanda bir hukukçunun hayatını ele alacağını okuyucuya iletmek
D) Romandaki bir karakterin geçmişi hakkında bilgi vermek

REFERENCES

Afflerbach, P. (2011). *Understanding and using reading assessment, K–12*. Newark, DE: International Reading Association.

Ahmed, Y. (2011). *Developmental relations between reading and writing at the word, sentence and text levels: A latent change score analysis* (Master's thesis). Retrieved from Electronic theses, treatises and dissertations. (Paper 4683)

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

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

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bernhardt, E. B. (1991). *Reading development in a second language: Theoretical, empirical, and classroom perspectives*. Norwood, NJ: Ablex.

Bernhardt, E. B. (2011). *Understanding advanced second language reading*. New York and London: Routledge.

Bossers, B. (1991). On thresholds, ceilings, and short-circuits: The relation between L1 reading, L2 reading and L2 knowledge. In J. H. Julstijn, & J. F. Matters (Eds.), *AILA Review*, *8*, 45-60.

Brown, J. D. (2005). *Testing in Language Programs*. New York, NY: McGraw-Hill.

Carrell, P. L. (1987). Content and formal schemata in ESL reading. *TESOL Quarterly 21*, 461-481.

Chen, C. (2010). On reading test and its validity. *Asian Social Science, 6*, 192-194,

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

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology, 99* (2), 311- 325.

Devi, S. (2011). Careful versus expeditious reading: the case of the IELTS reading test. *Academic Research International 1*(3), 25-35.

Diakidoy, I. N. N., Stylianou, P., Karefillidou, C., & Papageorgiou, P.(2005). The relationship between listening and reading comprehension of different types of text at increasing grade levels. *Reading Psychology*, *26*(1), 55-80.

Downing, S. (2006). Twelve steps for effective test development. In S. Downing, & T. Haladyna, M. (Eds.), *Handbook of test development* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum Associates.

Eisterhold, J. (1990). Reading-writing connections: Toward a description for second language learners. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 88- 101). New York: NY: Cambridge University Press.

Farhady, H., & Hassamy, G. Z. (2005). Construct validity of L2 reading comprehension skills. *Iranian Journal of Applied Linguistics, 8*, 29-53.

Goodman, K. S. (2001). *On reading*. Portsmouth, NH: Heinemann.

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly, 25* (3), 375-406.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York, NY: Cambridge University Press.

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading*. London: Longman.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Hirai, A. (1999). The relationship between listening and reading rates of Japanese EFL learners. *Modern Language Journal, 83*, 367-384.

Hirvela, A. (2004). *Connecting reading and writing in second language writing instruction*. Ann Arbor, MI: University of Michigan Press.

Hsueh-chao, M. H., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*, 403-430.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*, 219–230.

Hughes, A. (2003). *Testing for language teachers*. New York, NY: Cambridge University Press.

Jaliehvand, M., & Moses, M. (2014). Influence of rhetorical pattern on improving EFL sudents' reading comprehension. *Journal of Studies in Social Sciences, 7,* 210-225.

Jiang, X. (2011). The role of first language literacy and second language proficiency in second language reading comprehension. *The Reading Matrix, 11*, 177-190.

Joshi, R. M., & Aaron, P. G. (2000). The component model of reading: Simple view of reading made a little more complex. *Reading Psychology, 21*, 85-97.

Katalayi, G. B., & Sivasubramaniam, S. (2013). Careful reading versus expeditious reading: Investigating the construct validity of a multiple-choice reading test. *Theory and Practice in Language Studies, 3*, 877-884.

Kintsch, W., & Rawson, K. A. (2005). Comprehension. In M. J. Snowling, & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 209-226). Malden, MA: Blackwell.

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.

Koda, K. (2005). *Insights into second language reading*. New York, NY: Cambridge University Press.

Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development. In K. Koda (Ed.), *Reading and language learning* (pp. 1-44). *Special issue of Language Learning Supplement, 57,* 1-44.

Kumar, R. (2012). *Research methodology*. London: Sage Publications.

Lee, J., & Schallert, D. L. (1997). The relative contribution of L2 language proficiency and L1 reading ability to L2 reading performance: A test of the threshold hypothesis in an EFL context. *TESOL Quarterly, 31*, 713-739.

Leeser, M. J. (2007). Learner-based factors in L2 reading comprehension and processing grammatical form: Topic familiarity and working memory. *Language Learning, 57*, 229-270.

Liao, C. W., Qu, Y., & Morgan, R. (2010). The relationships of test scores measured by the TOEIC Listening and Reading Test and TOEIC Speaking and Writing Tests. In D. E. Powers (Ed.), *The research foundation for TOEIC: A compendium of studies* (pp. 13.1–13.15). Princeton, NJ: Educational Testing Service.

McNeil, L. (2011). Investigating the contributions of background knowledge and reading comprehension strategies to L2 reading comprehension: An exploratory study. *Reading and Writing, 24*, 883-902.

Messick, S. (1987). *Validity.* ETS Research Report no. 87-40. Princeton, N.J: Educational Testing Service.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1990). Validity of test interpretation and use. Princeton, N.J: Educational Testing Service. ETS Research Report no. 90-11.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Milanovic, M. (2002). *Common European Framework of Reference for Languages: Learning, teaching, assessment: Language examining and test development*. Strasbourg: Council of Europe, Language Policy Division. Retrieved from http://www.coe.int/T/DG4/Portfolio/documents/Guide%20October%202002%20revised%20version1.do

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*(4), 379-416.

Morvay, G. (2012). The relationship between syntactic knowledge and reading comprehension in EFL learners. *Studies in Second Language Learning and Teaching, 2*, 415-438.

Nassaji, H. (2003). Higher-level and lower-level text processing skills in advanced ESL reading comprehension. *The Modern Language Journal, 87*, 261-276.

Nunan, D. (1991). *Language teaching methodology*. Hertfordshire: Prentice Hall International.

Oakhill, J. V., Cain, K., & Bryant, P. E. (2003). The dissociation of word reading and text comprehension: Evidence from component skills. *Language and Cognitive Processes, 18*, 443-468.

OECD (2010). *PISA 2009 results: Executive Summary*. Paris: OECD.

OECD (2013). *PISA 2012 Results in Focus: What 15-year-olds know and what they can do with what they know*. Paris: OECD.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling, & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227-247). Oxford: Blackwell.

Rayner, K., Pollatsek, A., Ashby, J., & Clifton, C. (2012). *The psychology of reading*. New York, NY: Psychology Press.

Rozimela, Y. (2014). The students' genre awareness and their reading comprehension of different text types. *International Journal of Asian Social Science, 4*, 460-469.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning 52*(3), 513–536.

Sabatin, I. M. (2013). The effect of cultural background knowledge on learning English language. *International Journal of Science Culture and Sport, 1*(4), 22-32.

Sabatini, J. P., Bruce, K., & Steinberg (2013). *SARA reading components tests, RISE form: Test design and technical adequacy* (Research Report No. RR-13-08). Princeton, NJ: Educational Testing Service.

Sirin, S. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research 75*, 417-53.

Smith, F. (2004). *Understanding reading*. Mahwah, NJ: Lawrence Erlbaum.

Stanovich, K. E. (1980). Toward an interactive compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16*, 32-71.

Urquhart, S., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice*. New York: Longman.

Ünaldı, A. (2010). *Investigating reading for academic purposes: sentence, text, and multiple texts* (Unpublished doctoral dissertation). University of Bedforshire, Bedforshire, England.

Wang, H. (2008). Probing EFL Students' Language Skill Development in Tertiary Classrooms. *English Language Teaching 1*(2), 3-7.

Wang, J. H., & Guthrie, J. T. (2004). Modeling the effects of intrinsic motivation, extrinsic motivation, amount of reading, and past reading achievement on text comprehension between U.S. and Chinese students. *Reading Research Quarterly, 39*, 162–186.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke, UK: Palgrave Macmillan.

Weir, C. J., & Porter, D. (1994). The Multi-divisible or unitary nature of reading: The language tester between Scylla and Charybdis. *Reading in a Foreign Language*, *10*(2), 1-19.

Weir, C. J., & Khalifa, H. (2008). A cognitive processing approach towards defining reading comprehension, *Cambridge ESOL: Research Notes*, *31*, 2-10.

Wise, J. C., Sevcik, R. A., Morris, R. D., Lovett, M. W., & Wolf, M. (2007). The relationship among receptive and expressive vocabulary, listening comprehension, pre-reading skills, word identification skills, and reading comprehension by children with reading disabilities. *Journal of Speech, Language, and Hearing Research*, *50*(4), 1093-1109.

Zarei, A. A., & Neya, S. S. (2014). The effect of vocabulary, syntax, and discourse-oriented activities on short and long-term L2 reading comprehension. *International Journal of Language & Linguistics, 1*, 30-39.

Zhang, X. (2008). The effects of formal schema on reading comprehension: An experiment with Chinese EFL readers. *Computational Linguistics and Chinese Language Processing, 13*(2), 197-214.

Zhang, D. (2012). Vocabulary and grammatical knowledge in L2 reading comprehension: A structural equation modeling study. *The Modern Language Journal*, *96*, 554-571.

Zhou, L. (2011). Effects of text types on advanced EFL learners' reading comprehension. *Journal of Language and Culture, 30*(2), 45–56.