VALIDATING ASPECTS OF A READING TEST

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF SOCIAL SCIENCES

OF

MIDDLE EAST TECHNICAL UNIVERSITY

BY

ZEYNEP AKŞİT

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR

THE DEGREE OF DOCTOR OF PHILOSOPHY

IN

THE DEPARTMENT OF ENGLISH LANGUAGE EDUCATION

SEPTEMBER 2018

Approval of the Graduate School of Social Sciences

____

Prof. Dr. Tülin Gençöz
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Doctor of Philosophy.

____

Assoc. Prof. Dr. Bilal Kırkıcı
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy.

____

Assoc.Prof.Dr. Çiler Hatipoğlu
Supervisor

Examining Committee Members

Prof. Dr. Dinçay Köksal	(COMU, FLE)	________________________
Assoc. Prof. Dr. Çiler Hatipoğlu	(METU, FLE)	________________________
Assist. Prof. Dr. Necmi Akşit	(Bilkent Uni, FE)	________________________
Prof. Dr. Ayşegül Daloğlu	(METU, FLE)	________________________
Prof. Dr. İsmail Hakkı Erten	(Hacettepe Uni, ELT)	________________________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Last Name, Name: Akşit, Zeynep

Signature :

ABSTRACT

VALIDATING ASPECTS OF A READING TEST

Akşit, Zeynep

Ph.D., Department of English Language Education

Supervisor : Assoc. Prof. Çiler Hatipoğlu

September, 2018, 331 pages

This study investigated three aspects of validity (i.e., context, cognitive and scoring) of a reading test: First, the reading test construct was defined based on the cognitive processing model and its criterial features were presented in the test specifications. Secondly, to establish cognitive validity, the cognitive processes that were activated during test taking were investigated through retrospective and introspective verbal data. The results revealed whether the test elicited behavior reflective of reading activities beyond the testing situation. Finally, the scoring validity of the test was examined through item analysis. The findings revealed that the performances were generalizable: the tasks elicited behavior similar to those in real-life reading, and that the criterial parameters of the test were at an acceptable level for the intended population. Moreover, the majority of the item parameters were within the expected ranges.

This study and its findings have important implications. At the organizational level, the results have implications for improvement in testing practice: the sociocognitive framework provides a systematic approach and encourages the generation of evidence for validity at all stages of test development. Moreover, test scores that reliably reflect test takers' ability on relevant aspects of reading help improve reading instruction. Instruction that is grounded in theory and supported with established needs can help students be equipped with skills required for successful academic reading. Implications at the theoretical level are: The cognitive processing reading model, which was originally based on L1 reading, successfully defines L2 academic reading within the context of this study.

Keywords: reading test, context validity, cognitive validity, scoring validity, cognitive processing model of reading

ÖZ

BİR OKUMA SINAVININ BAZI YÖNLERDEN GEÇERLEMESİ

Akşit, Zeynep

Doktora, İngiliz Dili Öğretimi Bölümü

Danışman : Doç. Dr. Çiler Hatipoğlu

Eylül, 2018, 331 sayfa

Bu çalışma, tamamlayıcı karma yöntemlerle bir okuma sınavının bağlam, bilişsel ve notlama geçerliğini araştırmıştır. Öncelikle, okuma sınavının kurgusu bilişsel süreç modeli kullanılarak tanımlanmış ve kurgunun kriterleri sınav tanımlamaları dosyasında sunulmuştur. İkinci olarak, sınavın bilişsel geçerliği için sınav sorularına cevap verme sırasında etkinleşen bilişsel süreçler geçmişe dönük anımsama ve sınav anında içebakış yöntemleri ile incelenmiştir. Verilerin analizi ile sınav sorularının sınav dışındaki ortamlarda uygulanan okuma biçimlerini yansıtıp yansıtmadığına karar verilmiştir. Son olarak, notlama geçerliği madde analizleri ile incelenmiştir. Bulgular, sınananların sınav performansının hedef ortam için genelleştirilebileceğini göstermiştir: Sınav soruları gerçek hayatta okumaya benzer süreçlerin etkinleştirilmesini gerektirmiştir. Ayrıca, sınav sorularının kriterlerinin hedef kitle için uygun düzeyde olduğu anlaşılmıştır. Son olarak, notlama geçerliği açısından, maddelerin çoğunun parametreleri yeterli bulunmuştur.

Bu çalışma ve bulgularının önemli çıkarımları vardır. Kurumsal düzeyde sonuçların sınama uygulamalarını iyileştirmeye yönelik çıkarımları vardır: sosyal bilişsel çerçeve sınav geliştirmenin her aşamasında sistematik bir yöntem sunmakta ve geçerlik kanıtı oluşturulmasını desteklemektedir. Ayrıca, sınav sonuçları sınananların okuma becerisini güvenilir bir şekilde yansıttığından, okuma becerisinin öğretiminde de olumlu etkisi olacaktır. Kuramsal altyapısı güçlü, ve araştırmalar ile belirlenmiş ihtiyaçlara cevap veren bir eğitim sistemi öğrencilerin akademik okuma için gerekli becerileri elde etmesine yardımcı olur. Sonuçların kuramsal düzeyde çıkarımı da vardır: anadilde okuma için hazırlanan bir bilişsel süreç okuma modeli, ikinci dilde okumayı tanımlamada ve okuma becerisini sınamada başarıyla kullanılmıştır.

Anahtar kelimeler: okuma sınavı, bağlam geçerliği, bilişsel geçerlik, notlama geçerliği, bilişsel süreç okuma modeli

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Associate Professor Çiler Hatipoğlu, for her continuous support of my Ph.D study, and advice she has provided throughout the writing of this dissertation.

I would like to extend my gratitude to my dissertation committee member Professor Ayşegül Daloğlu, for her warm encouragement and constructive criticism and, other members of my dissertation committee, Professor Dinçay Köksal, Associate Professor Necmi Akşit, and Professor İsmail Hakkı Erten for their insightful comments.

I am grateful to Associate Professor Aylin Ünaldı, who provided guidance during the initial stages of my research, and meticulously read the first chapters of my dissertation.

I greatly appreciate the Proficiency Committee members, Işık Arıkan, Özlem Polat, Şükran Saygı and Dr. Vildan Şahin, for taking me in as a teammate and supporting me throughout the writing of this dissertation. Without their precious support it would not have been possible to conduct this research. I extend special thanks to Şükran Saygı, who has proofread parts of my dissertation, and has always been a great help whenever I needed.

Many thanks also to Özlem Atalay and Naz Dino for their continuous support, Professor Hüsnü Enginarlar, for introducing me to the challenges of testing, Professor Lyle Bachman, for reading the abstract of my dissertation and providing valuable feedback, and Joe Hobbs, for reading my proposal and teaching me to own my work.

I would also like to say a heartfelt thank you to my parents, Ruhiye Kayra and Muhsin Gösterişli, for supporting me throughout all my studies, and Zerrin and Sinan Tandoğan for their helpful comments and support.

Finally, I have to thank my husband, Mehmet Akşit, to whom I am greatly indebted for his guidance during my dissertation study, for his unwavering support, belief in my abilities, and invaluable insight into research, and my children, Defne and Ateş, for their love and support, and patience that cannot be underestimated.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Educational institutions employ measures of language knowledge widely and for a number of reasons. One major function of language assessment in higher education is gatekeeping: Schools make admission or certification decisions based on test results. In admissions, test results are used to decide whether the applicants' knowledge of or ability to use the language of instruction is sufficient for them to meet the requirements of academic studies. Similarly, certification decisions relate to whether the applicant is able to perform certain operations at a desired level. To this end, various types of language assessment batteries are used. Depending on the context, either it can be a general proficiency test developed by an external institution or an in-house proficiency test developed with a concern for examining needs relevant in a particular context.

Proficiency tests are tests that claim to measure general language ability. There are international organizations that develop such tests for different contexts and for people with different training backgrounds (Alderson, Clapham, & Wall, 1995). In-house tests, on the other hand, are generally based on the established needs of academic life in the relevant contexts. Both types of tests are called high-stakes tests as the scores from such tests are used to make high-stakes decisions. For instance, failure on such a test may result in refusal to or dismissal from a program. Therefore, in high-stakes testing, as well as in other testing situations, it is the responsibility of test developers to ensure that the assessment battery serves its purpose fairly and meaningfully. To this end, one of the main concepts that dominate both the pre– and post–operations concerning the design, development and administration of a high-stakes test is validity.

Validity is about the usefulness, fairness and meaningfulness of a test (Messick, 1989b). In earlier definitions, validity was considered as a quality of a test, and a test was considered to be valid if it measured what it claimed to measure. However, with the advances in the field in the 1950s, the definition and the approach towards validity was radically modified. Today, rather than being a quality of a test, validity is judged by the extent of the evidence provided by test developers to confirm that score-based decisions are justifiable (Chapelle, 1998; Council of Europe, 2009; Cronbach, 1988; Fulcher & Davidson, 2007; Kane, 2016; McNamara, 2006; Messick, 1995; Mislevy, 2007; Moss, 2007; Read & Chapelle, 2001; Sireci, 2009; Weir, 2005a). Admission or certification decisions need to be backed up with research as warrants of their validity. Only then can the decision makers ensure that their decisions of admission, or refusal based on test scores are just and meaningful.

Many organizations or institutions who develop and administer standardized tests carry out extensive research to be able to back their claims about the meaningfulness, appropriateness or fairness of their exams. ETS is one of the well-known educational testing and assessment organization who develops and administers standardized tests, one of which is the renowned TOEFL. Another one is the Cambridge English who develop English exams at different levels. The line of research regarding validity by these organizations includes studies on construct representation (Biber & Gray, 2014; A. D. Cohen & Upton, 2006), on authenticity and content validity (Rosenfeld, Leung, & Oltman, 2001; Stricker & Attali, 2014), on criterion-related and predictive validity (Weigle, 2014), on validation research on tests of discrete skills (Cartwright, 2009; Shaw & Weir, 2008; Taylor, 2011), and validation methods (Grotjahn, 1986; McNamara, 2006; Wilson, 1999).

This study is an attempt to generate validity evidence for a reading test. More specifically, it aims to generate evidence for contextual, cognitive and scoring validities of reading text. Khalifa and Weir (2009) posit that these three validities constitute construct validity, which is the central concept in validity theory.

1.1 Background of the Study

Middle East Technical University (METU) is described as an international research university on its website (www.metu.edu.tr). The international character of the university is achieved through its partnerships with international institutions, the funds generated from international research projects, and accommodation of international researchers and students.

The medium of instruction at METU is English. Hence, all candidates wishing to study at METU are required to provide proof of a certain level of English language proficiency before they can start their academic studies. The School of Foreign Languages (SFL) at METU administers a test, English Proficiency Exam (METU-EPE), to the newly registered students at the onset of each academic year. In general, the minimum score that test takers need to be able to move to any of the undergraduate programs at METU is 60 over 100, except for the Foreign Language Education department, which requires 70. Alternatively, equivalent scores from language examinations given by one of the two external organizations; namely, the TOEFL IBT (75 – 86 points) by Educational Testing Service (ETS) (www.ets.org) and IELTS (6.0 – 7.0) which is jointly owned by British Council, IDP: IELTS Australia and Cambridge English Language Assessment (www.ielts.org) are also accepted as valid proof of English language proficiency.

The students who obtain the required scores from METU-EPE start their subject studies whereas those who fail to receive the required minimum score, study at the Department of Basic English (DBE) for one year before taking the METU-EPE again at the end of the instructional period in June. Instruction at DBE is focused on providing foundational English to students who lack the necessary language skills and prepare them for academic studies.

METU-EPE is given four times every year. Here is the schedule and the test taking cohort:

- September: newly registered undergraduate students and the graduate candidates,
- December: graduate candidates
- June: DBE students (except for those students who start English education at the beginner level) and graduate candidates

- August: DBE students who started English education at the beginner level in the fall semester (pre-intermediate level group)

Documentation related to METU-EPE are as follows: for the candidates, there is a web page on METU servers (www.metu.edu.tr) and a booklet on sale at METU bookstore. The information in both media consists of the format of the exam, the sections and the question types, sample questions and the scoring rubric. This type of information helps the candidates understand what is expected of them, so that they can prepare better and perform better.

Other documentation on the exam concerns the instructors: some statistical information about the exam is provided to the instructors at the DBE and MLD. The former director of the SFL used to announce some of the results of the score-related analysis, such as correlation coefficients and averages. Some instructors are known to have carried out small scale research on METU-EPE, and there is one unpublished Master's thesis (Ataman, 1999) on the validity of an earlier version of the proficiency exam. There is no other official document or published research on the design and development of METU-EPE to the best of my knowledge.

This lack of documentation on METU-EPE casts doubt on the fairness and meaningfulness of the decisions given by the registrar's office at METU, and points to an urgent need for more systematic research on aspects of the test such as the content, the theoretical underpinnings, or the essential criterial features of the test.

1.2 Statement of the Problem

In high stakes tests, the decisions taken or inferences made about the test takers have important consequences, and erroneous decisions cannot be easily reversed and remedied (Bachman & Palmer, 1996). METU-EPE is a high-stakes test and the consequences of the decisions based on the scores of METU-EPE are grand. Newly registered undergraduate students make up the majority of the test takers whose scores are used to decide whether to allow them to start studying in an academic degree program, or to delay their academic study for a year (or sometimes two) while they attend the DBE to improve their language skills. In the case of graduate applicants, their scores from METU-EPE are used to decide whether to allow or deny admission to a graduate degree program. Since the test has a major impact on stakeholders, it is essential to justify the decisions taken based on the test results by following relevant validation procedures (Messick, 1989b).

Validation is the process of operationalizing the concept of validity (van der Walt & Steyn, 2008). In the 1940s, the early pragmatic and empirical view of validity was based on correlation and factor analysis (Sireci, 2009). However, a seminal article by Cronbach and Meehl (1955) radically changed the concept of and approach towards validity. It was no longer considered as a feature of a test but a unitary concept that reflected the intrinsic relations between various constituents of a test. This novel approach to validity called for the analysis of different aspects in a testing situation such as the content of the test, the interaction between the test taker and the test tasks, and the predictive power of the test. It also provided a framework to gather evidence about the validity of the test (Sireci, 2009). A second seminal article on validity appeared in 1989. In this article, Messick emphasized that validity claims are not about the test but about the interpretations of the scores of the test. And, in order to claim validity, empirical evidence as well as theoretical rationales that support the adequacy and appropriateness of inferences based on test scores are needed.

Acknowledging Messick's (1989b) approach to validity, it follows that the inferences made on the scores of any high-stakes test should be backed by empirical research and a sound theoretical underpinning. METU-EPE being a high-stakes test, the test developers have the responsibility to gather and present evidence that explains how and why the test scores are valid and reliable indicators of the ability that is assessed with that specific instrument. A similar claim was made by Chalhoub-Deville (1997) who said

> … in high stakes testing where critical decisions are made (e.g. certification, fulfilling a degree requirement, admission into a programme, progressing into a higher grade, securing a job, etc.), it is imperative that resources be allocated for assessment frameworks to be validated in their context of use. In high-stakes testing, the deficiency of evidence to support an assessment framework in a given context of application weakens the validity of test interpretation and use, which has grave ramifications. (p. 17)

In the case of METU-EPE, nothing much is known about the theoretical basis of the test, or rationale for the score interpretations such as "if a test taker receives 85/100 in the test, she will be exempt from the first year English courses". Neither are there any documents specifying the objectives of the exam, or guiding principles for item/test development. This is not surprising considering that it was not a common practice to produce exam specifications in the past (Weir, Huizhong, & Yan, 2000). "The construct of reading that is measured in the TOEFL reading test is not made explicit in the ETS literature" (Peirce, 1992, p. 668), for instance. In the case of METU-EPE, the only available formal document related to the test was a booklet prepared for the test takers that included introductory information about the different sections of the test, samples of different item types and scoring rubrics.

Considering the impact of score-based inferences on critical decisions regarding student admission to academic degree programs and the impact of METU-EPE on instruction at the DBE and the MLD, it is clear that principled research on various stages of test development is necessary to ensure that the inferences made from test scores are meaningful, appropriate and useful (Messick, 1989b). Notwithstanding research studies by international testing organizations (for example, Educational Testing Service – ETS, Cambridge ESOL) and local universities (for example, Bilkent University, Boğaziçi University) where in-house proficiency tests are developed, it is vital to carry out validation studies in own/specific contexts because contextual variables affect many aspects of validity. As Brown and Goodman maintain "Validity claims always occur in and are tied to specific contexts" (2001, p. 206). Hence, there is a need to investigate and report test design and development stages in METU context for accountability to the stakeholders.

Validation of a test is essentially combining the theoretical rationale with empirical research to show that the interpretations based on test scores are justified. Cohen (2006) posits that empirical research on test-taking strategies is necessary if we want to understand what tests actually measure and to make sure that performance on a test can be generalizable; that is, test taker's performance is reflective of the expected behavior in the target language use domain, that is the "specific setting outside the test itself that requires the test taker to perform language use tasks" (Bachman & Palmer, 1996, p. 44). Currently, there are two main approaches used in test validation:

> (i) the argument-based approach that is concerned with developing and evaluating interpretive arguments by analyzing various types of theoretical and empirical evidence (see, for example, Bachman & Palmer, 2013; Kane, 1992; Mislevy, 2007),
>
> (ii) the evidence-based approach that views test validation as "… the process of generating evidence to support the well- foundedness of inferences concerning trait from test scores …" (Weir, 2005b, p. 1) and accumulating this evidence before and after test events.

This study utilized Weir's evidence-based approach to validation. From the point of Weir's (2005b) sociocognitive framework, *a priori* validity evidence is needed such as a blueprint or test specifications in writing which provides guidance to the test writers in item writing, and guidance to school administration in test administration and scoring procedures. In the case of METU-EPE, test and item preparation practice is based on experience, as there is no document that specifies item writing rules and procedures. Item writers use their own judgments to develop new items similar to the ones used previously. Decisions about the content of the test such as task types, topics, or difficulty levels of texts were most probably taken in the past; however, again, there is no document that reveals whether any theoretical or empirical study was carried out to justify these decisions. Similarly, about the scoring procedures, the grounds for current practice is unknown. In terms of *a posteriori* validity evidence, a number of statistical analysis within the Classical Test Theory (CTT) is carried out after each administration of the test and reported to the SFL administration and METU-EPE item writers.

While attempts have been made to analyze and interpret statistical outcome of the test scores, a thorough investigation of the test with regard to content, construct, and scoring validities are missing. It follows that in its present state, it is difficult to make sound generalizations about test takers' ability to use language as required in academic programs at METU.

In an attempt to fill in this gap, this study aims to validate the reading section of the English proficiency test that was developed as part of the program evaluation project at the SFL. It specifically investigates the theoretical basis of the test, the content, and the properties of the test items to generate evidence enabling the justification of the decisions made about the test takers' reading ability in the context of first year academic degree programs at METU.

1.3 Research Questions

This research study on validation of aspects of the reading test was carried out in three stages: in the first stage test development procedures were carried out (as part of *a priori validation*), in the second and third stages cognitive validity (again part of *a priori validation)* and scoring validity (as part of *a posteriori* validation) of the test was investigated.

A priori validation is about defining the abilities that are relevant to the testing context, both theoretically and operationally. In order to arrive at a viable definition, both the theoretical literature and research literature were reviewed. As for the operational definition of reading, a needs analysis study that had been carried out in the target language use context previously was reviewed meticulously. Furthermore, the literature was reviewed for other studies that dealt with the analysis of the real life tasks in the target language domain in an academic environment. Combining the information from these sources a model was proposed and a pilot test was developed operationalizing the reading construct.

A posteriori validation, on the other hand, is mainly concerned with analyzing the test data to establish that item statistics support the interpretations of the test scores. Other outcomes of *a posteriori* validation are evidence for concurrent and consequential validities of a test, both of which were left outside the scope of this study.

Hence, the first and second research questions deal with the conceptualization and operationalization of the reading ability, questioning context and cognitive validities of the test (*a priori* validity) whereas the third research question is focused on the scoring validity (*a posteriori* validity):

1. How is the academic reading ability conceptualized and operationalized as a test construct?
2. What are the cognitive processes that underlie the construct of the reading test a) in retrospection and b) in introspection?
3. To what extent do item parameters contribute to the validity claims of the test?

1.4 Significance of the Study

In many state and private Turkish universities, English is used as the medium of instruction. Hence, it is a common practice in such institutions to use various assessment instruments to make admission decisions regarding language proficiency. Some language schools chose to administer tests that are prepared by international organizations such as the Educational Testing Service's (ETS) TOEFL or the IELTS that is a product of a collaboration between British Council and IDP Australia. Examples of such institutions that use external tests are Koç University in Istanbul and TOBB ETÜ University in Ankara. Others prepare their own English language tests. Some of the public (P) and foundation (F) universities that develop their own proficiency tests are: Ankara Yıldırım Beyazıt University (P), Atılım University (F), Bahçeşehir University (F), Başkent University (F), Bilgi University (F), Bilkent University (F), Boğaziçi University (P), Çağ University (F), Erciyes University (P), Gazi University (P), Hacettepe University (P), Sabancı University (F), and TED University (F). In places where local tests are used, test development process is usually regarded as a knowledge-base that is proprietary information and the specifics of this process are usually kept confidential. The amount of information from within those institutions is limited[1].

[1] Some of the published work and unpublished thesis/dissertation on test development projects in Turkey are: a monograph on the development of the Bosphorus University English Proficiency Test (BUEPT) by Arthur Hughes, who directed English language testing project at Boğaziçi University in the years 1982 -84, and a doctoral dissertation on the reading test of BUEPT by Aylin Ünaldi (2004). According to the Thesis Center of the Council of Higher Education (https://tez.yok.gov.tr/UlusalTezMerkezi) there are five other studies (MA Theses)

At the local level, then, this study will improve practice in the field of language teaching and testing by presenting how an assessment and validation framework is implemented in an academic English as a Foreign Language (EFL) context to develop a high-stakes language test. Test developers, item writers, instructors, registrars' officers, administrators, and other policy makers will be informed of the procedures and processes carried out in test development and validation. Awareness of good practice in assessment may help to eliminate testing habits carried out intuitively or test development approaches that are not grounded in any theory of language, and thus cannot be considered valid or reliable measures. This awareness may also help to give more informed decisions on the meaning and generalizability of test scores.

Furthermore, this research made use of the results of a needs analysis project that had been carried out at METU, and also reviewed the literature to compare and contrast analysis of reading tasks relevant in similar contexts. Hence, it will provide sound basis for test development in similar EFL contexts by providing information on how to merge local needs with a theory to produce an exclusive working model for their own context.

This research will also pave the way for a more systematic approach in test development. Testing as a field of study has limited popularity as the number and content of courses on testing are inadequate (Hatipoğlu, 2015). The number of people with formal training on test development being quite small, test developers or item writers are usually chosen from among experienced instructors whose experience in teaching and background knowledge on the test taking populations are believed to be advantageous in producing language tests. Whereas expertise in teaching is a very important asset in developing tests, it is not sufficient. Assessment literacy, in the sense of knowledge of test development and validation procedures, is as important as the knowledge of content matter. In fact, it is considered as the *sine qua non* condition for educators (Popham, 2009). As such, this study might inflict interest in testing as an

on the validity of English proficiency tests (Ataman, 1999; Gürsoy, 2013; Kutevu, 2001; Yapar, 2003; Yeğin, 2003). However, those studies focus on the product of the tests (test scores) rather than the test development process.

important component in the field of language teaching, and create awareness about different approaches and procedures in test development.

In high-stakes testing, different from classroom testing, test developers are accountable for presenting validity evidence of the test (Hatipoğlu, 2016). Hence, there is a need to invest time and energy both to develop tests and to validate the decisions and inferences made on test results. This study contributes to the field of high-stakes testing by presenting the stages of test development and the findings of every stage in a meaningful and transparent manner.

At the global level, this study contributes to the wider knowledge base of the application of a framework for validation purposes. Utilizing an assessment/test development framework provides a sound basis on which to build an assessment instrument. In addition, as the framework is used in a wider variety of contexts and in different backgrounds, it is possible to acquire more information on the different facets of testing and whether all aspects presented in the framework are viable in contexts other than it was developed. This study, hence, helps to generate a new conceptualization of the reading ability in an EFL context using tools developed in a non-EFL context. Test validation carried out in the present study prioritizes generation of validity evidence in accordance with the contextual parameters, local needs, and test-taker characteristics, which in turn have an impact on how the reading construct is conceptualized and operationalized appropriate to the setting of the assessment.

CHAPTER 2

LITERATURE REVIEW

All tests yield a result. In language testing it is either a numerical score or a letter grade. In any case, the test administrators, or policy makers make a decision or draw a conclusion about a test taker's ability/knowledge/skill from these results, which is called a score-based inference. It is basically a prediction about a test taker's future performance on a specific task in real life.

In the previous chapter, it has been claimed that a systematic approach in assessment is needed to be able to draw meaningful and reliable conclusions about test-takers. This systematic approach is achieved through the use of an assessment/validation framework. This study utilizes a framework in the validation of a reading test and defines reading ability through the use of a model of reading proposed by Khalifa and Weir (2008a). This framework recommends methods to investigate the context, cognitive and scoring validities of a reading test.

In the following sections, I present a review of the concept of validity to shed light on the rationale of the validation framework of this study. I also present some studies which utilized frameworks and models in test development and validation.

2.1 Historical Development of the Concept of Validity

In any research study on a testing situation there is reference to validity. Chapelle (1999) says,

> [T]he definition of validity affects all language test users because accepted practices of test validation are critical to decisions about what constitutes a good language test for a particular situation. In other words, assumptions about validity and the process of validation underlie assertions about the value of a particular type of test. (p. 254).

As the value of a test is closely related to the definition of validity for that particular testing situation, it is essential to understand the concept well and know how to extract the necessary information from the test and the scores. I will try to unveil the overwhelming and complicated nature of validity by reiterating the evolution of the concept starting from the early 20th century.

The modern concept of validity first emerged in early psychometric literature in the 1920s and it was essentially a pragmatic approach that viewed any correlation of the test as validity indicator (Sireci, 2009). At around the same period, Spearmen developed factor analysis, which became a popular tool to unveil the traits that underlie the performance of the test takers. Guilford (1946) was one of the proponents of using factor analysis in establishing the validity of a test and he categorized validity as 1) *factorial validity* referring to factor loadings of the test on meaningful factors and 2) *practical validity* referring to correlations between test scores and relevant criteria. Starting with correlation studies and followed by factor analysis, early views on validity were pragmatic and empirical. A theoretical definition was also proposed at that time which claimed that validity is the extent to which a test measures what it is supposed to measure (Garrett, 1937).

Towards the middle of the 20th century, some discontentment among psychometricians surfaced on the basis that test validation was limited to some statistical procedures, namely correlation and factor analysis, and that the criteria against which the tests were correlated were not sufficiently defined (Jenkins, 1946). Among a number of reasons for the unreliable nature of the criteria, Jenkins said that they may not be valid because of the "failure of the criterion-measure to comprise a large and significant part of the total field of performance desired" (Jenkins, 1946, p. 95). What followed was the apparent need to examine carefully the attributes that were the focus of the measurement, how these attributes were to be defined operationally, and the analysis of test content. This analysis was to demonstrate a sound relation between the procedures used in the criteria measure and the interpretation and stipulated use of the scores (Sireci, 2009). Using a sample of some performance as an indicator of a level of the target skill or ability, one could argue for the validity of the interpretations made of the examinee (Kane, Crooks, & Cohen, 1999). This emphasis on both the content of the test and the criteria was a novel perspective which was later called content validity.

The questions regarding how to treat validity and how to validate tests received attention from many psychometricians and theorists including Rulon (1946, in Sireci, 2009) who said that it was not possible to say whether a test is valid without defining its purpose and that different kinds of evidence was needed for validity. Soon, a committee was set up by the American Psychological Association (APA) to formally define test standards: how to construct, use and interpret tests. The committee announced four categories of validity:

1. Predictive validity: how well a test predicts performance on an external criterion,

2. Status validity: later named concurrent validity, which is concerned with the relationship between what is measured by a test and another existing criterion measure

3. Content validity: specifying the domain that is sampled for testing, and

4. Congruent validity: later named construct validity, which is about the quality of a test in terms of the theoretical model on which it is based.

These four categories were later reduced to three; namely, content, criterion-related and construct validities. A few years later, the concept of construct validity was elaborated in an article by two of the members of the APA committee: Cronbach and Meehl (1955). They stated that construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not operationally defined. The problem faced by the investigator is 'What constructs account for variance in test performance?' (p.282).

Cronbach and Meelh (1955) believed that the notion of construct validity was appropriate for psychological tests rather than educational. They said, "construct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured" (p.282). It was an alternative to the criterion and content models (Kane, 2001, 2012). However, soon, construct validity was found to be applicable to all educational tests as well as psychological. Both Loevinger (1957) and later Messick (1989b) argued that construct validity should be sought as it is not possible to define criteria of content, universally.

In his view of validity, Messick (1995) integrated "considerations of content, criteria and consequences into a comprehensive framework for empirically testing rational hypotheses about score meaning and utility" (p.742). This validity framework was first published in a seminal article in 1989, which seriously changed the way validity was approached. In this article, he introduced a unified view of validity and described it as an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p.13).

In this view, validity is not a property of a test but it is the extent to which we are justified in making inferences or giving decisions based on the test score. Hence, he proposed a progressive matrix with different facets as sources of evidence that contributed to this unified view of validity (Table 1). It is called progressive because in each cell there is construct validity but an additional facet is added starting from upper left cell.

Table 1 Messick's validity matrix (1989a)

	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

The evidential basis of *test interpretation* is construct validity. Messick (1989a) posited that construct validity can be achieved through evidence and rationales that provide proof about the *trustworthiness* of the meaning attributed to test scores. The evidential basis of *test use* is again construct validity; however, there is reference to relevance and utility in this dimension. It means, it is necessary to provide evidence that 1) the scores are relevant to the purpose of the assessment, and that 2) the scores have utility in the context where they are applied.

The consequential basis of *test interpretation* is related to the theory and philosophy that underlies the test. The consequential basis of *test use* is about the social consequences when test scores are used to make decisions about the test taker. This scheme provided guidelines for producing evidence of validity. Hence, test validation calls for 1) a hypothesis about the appropriateness of test outcomes, that is, test interpretation and test use, 2) data collection relevant to the hypothesis, 3) drawing a conclusion about the validity of test outcomes (Chapelle, 1999).

The type of data, i.e. evidence, that can be used in hypothesis testing were identified by Messick (1990). He suggested looking at test content with relation to the content of domain of reference

- examining the internal structure of test responses by checking the relationship between the responses in terms of tasks, items or parts,
- examining the external structure of a test by contrasting test scores with scores from other measures,
- examining how the different versions of the test brings about differences in responses, and
- controlling the social consequences of interpreting and using the test scores to understand what intended and unintended side effects occur.

Messick's view of validity was accepted in psychological, educational and language testing (Fulcher & Davidson, 2007). This is apparent in the consecutive revisions of the Standards for Educational and Psychological Testing: The four types of validity described in the Technical Recommendations (American Psychological Association, American Educational Research Association, 1954) were later revised to three types of content, criterion and construct validity in 1966. The 1999 Guidelines (Wilkinson, 1999) posited that there are no distinct types of validity since validity is a unitary concept. It further stated that validity "is the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests. [...] The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations." (p.9). This view of validity and validation posits a number of arguments:

1) what is validated is the interpretation of the scores rather than the scores or the test itself
2) there is a need to extensively analyze inferences and assumptions to interpret test scores and the interpretation will involve a rationale and will consider other possible interpretations
3) test users are expected to justify using a test score in a particular manner, and this justification involves demonstrating the preponderance of the positive consequences over the negative, and finally,
4) validation is a systematic effort in evaluating the interpretations of test scores rather than simply a collection of techniques: there should be consistency in the goals and approach to validation, and the criteria in judging the methods of validation (Kane, 2001).

Messick's conceptualization of validity greatly influenced research on test validation (Chapelle, 1999). A few examples from the current literature are Bachman (1990) who adopted Messick's validity concept to develop his argument-based approach in test validation; Shaw and Weir (2008) who consider context validity, cognitive validity and scoring validity as establishing construct validity of a test, and Weir (2005a) who defines reliability as part of scoring validity rather than being a distinct feature of a test. Nevertheless, the complicated character of the conceptualization of Messick's validity framework (O'Sullivan & Weir, 2011) gave rise to a more practical and operational description of validity (*Materials for the Guidance of Test Item Writers*, n.d.) in the form of test validation frameworks. Weir's sociocognitive framework is one where he presented a model that is used as a basis for test development as well as for validation

This overview of the historical development of the concept of validity and validation reflects the tension between the positivist and interpretivist approaches toward understanding the world around us (Davies & Elder, 2005). Correlation studies denote typical positivistic paradigm whereas the interpretivist paradigm views validity as “achieving consensus across multiple audiences and sources of evidence” (Lynch, 2003, p. 154). Within this interpretivist paradigm, it is expected to utilize a number of methods to elucidate different aspects of an issue. In the case of test validation, this approach points to the need to investigate characteristics of a test, or test facets, using a number of methods, procedures and techniques. In the present study, this was accomplished through the use of the socio-cognitive framework introduced by Weir (2005a), which suggests the use of theoretical perspectives as well as qualitative and quantitative methodologies in test validity inquiry.

2.2 Validation Through the Socio-cognitive Framework

Weir (2005a) asserts a general view of validity as,

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

Like Messick (1995), Weir emphasizes that test score is the means to establish validity, and that validity is multifaceted and different types of evidence are needed to support the validity claims of a test. However, he also argues that what is previously maintained as sources of evidence for different validities actually serve to establish construct validity which is a superordinate concept that embraces all forms of validity (Weir, 2005a). Nonetheless, the term *validity* has been used throughout his book to refer to this superordinate category, and reliability was considered as one form of validity evidence.

Weir’s (2005a) framework provides a criterial model for each of the four skills – reading, listening, writing and speaking – separately, and each model comprises five domains to generate evidence for the justification of the inferences made. The domains are: context validity, theory-based validity, scoring validity, consequential validity and criterion-related validity. With some variations in the criterial features for each skill, the generic model provided in Figure 1 clearly depicts the importance of the test taker on the interpretations of the consecutive operations, i.e. the procedures to generate validity evidence on the five domains.

The detailed version of the test development and validation model for reading is given in Figure 2. Weir (2005a) categorizes the procedures carried out to generate validity evidence before and after the test as *a priori* and *a posteriori* validation procedures, respectively. *A priori* validation refers to the procedures related to context validity and theory-based validity, and *a posteriori* validation refers to scoring, consequential and criterion-related studies on generating evidence for validity.

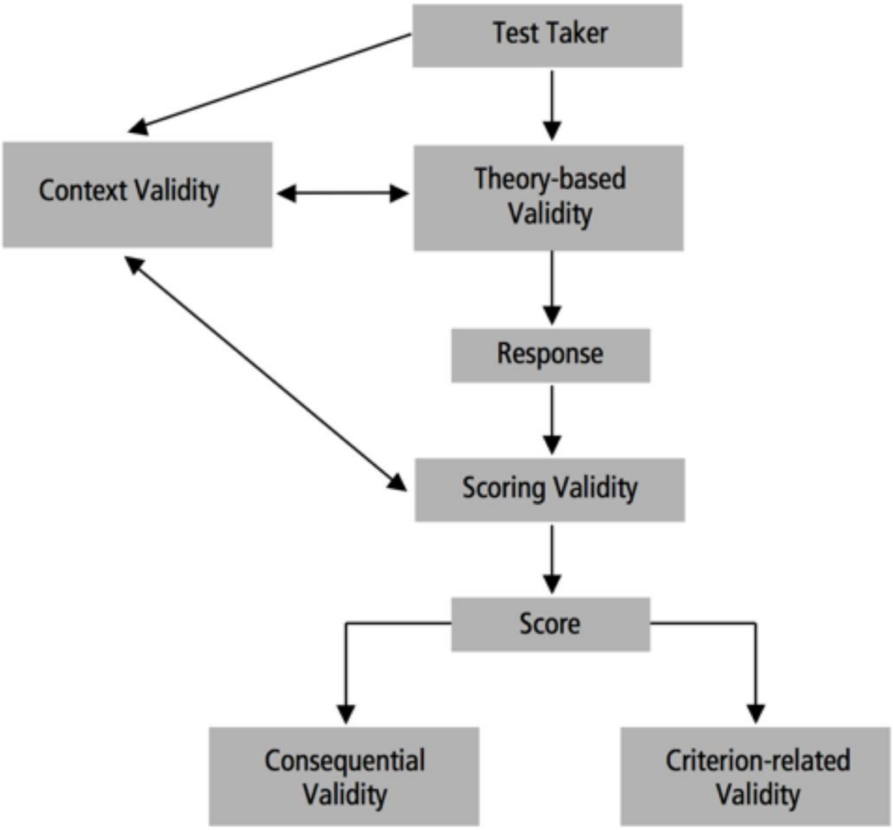


Figure 1 Weir's (2005) validation framework

Test taker characteristics, that is, test takers' physical/physiological, psychological and experiential characteristics have an impact on various aspects of the test, most importantly, theory-based validity and context validity. It has been argued that when designing tasks for a specific population, their characteristics, such as, age, sex, background knowledge, education, etc. need to be carefully evaluated so as not to create advantage or disadvantage for a specific group of people (O'Sullivan, 2000; Weir, 2005a).

2.2.1 Before the test events: Theory-based validity. During the structuralist period, in the 1960s, validity evidence was collected after the test event in the form of numerical data which were analyzed for factor loadings or correlation indices. However, Weir (2005a) maintains that it is problematic not to have a clear idea about the constructs of the test before administering it to the students. He makes a reference to Messick (1989b) who listed the two major threats to validity as construct-underrepresentation and construct-irrelevance. Weir (2005a) argues that we need to make sure before the test that the test construct is actually what we intend to measure and that there are no irrelevant variables. Otherwise, the test may have negative washback on instruction. For example, in construct-underrepresentation, if some important aspect of a skill is not tested, then it may not be taught either, as it is not included in the test.

Another argument for the use of theory-based validity is that theories of language models explicate the processes of language use. The test developer needs to demonstrate that the processes carried out during the real-life events are replicated in the test as closely as possible in order to claim theory-based validity. However, as reading processes are unobservable, we need to find a way to assess them through observable actions. Weir posits (2005a) that assessment of reading may follow a path similar to that in teaching: testing the component skills and strategies of reading. Hence, he suggests identifying skills and strategies that contribute mostly to the process of reading and assessing reading through them. Here, a *skill* is used to refer to automatized actions carried out usually subconsciously whereas a *strategy* refers to conscious problem solving activities (A. D. Cohen, 1998; Urquhart & Weir, 1998).

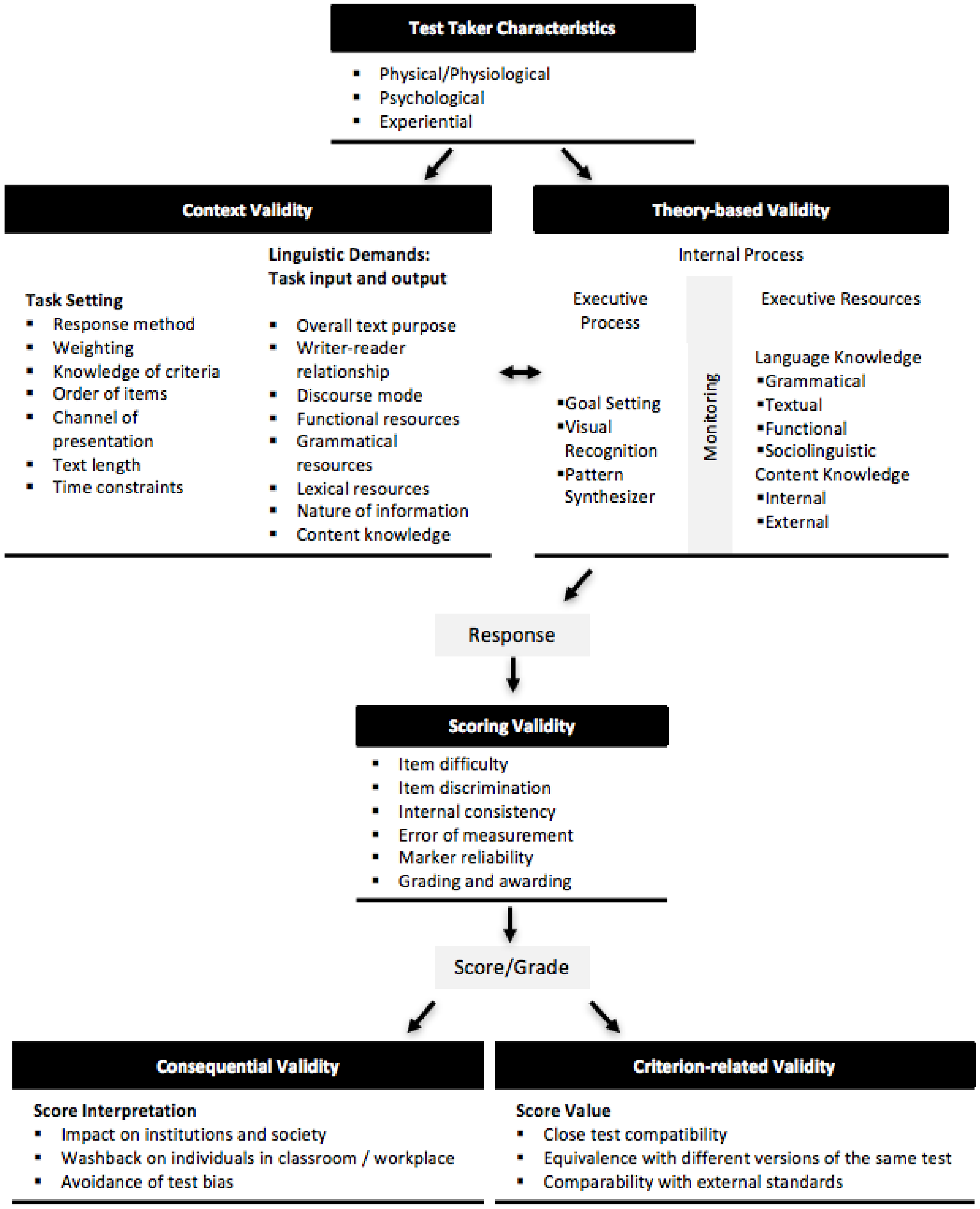


Figure 2 Aspects of validity for reading (Weir, 2005, Khalifa and Weir, 2009).

While suggesting to view reading as a componential skill for testing purposes, Urquhart and Weir (1998) added a second dimension into its assessment: reading at a local or global level. While local refers to comprehension of microstructures such as the lexical items, or references at the clause or sentence level, global reading refers to the comprehension at the macrostructure level; the main ideas, and any other important details. (The details of this model are given in Section 4.2.1)

The theoretical construct of reading discussed above was further elaborated in the theory-based validity component of the framework which now comprises of the subheadings executive processes and executive resources. Executive processes refer to setting a goal for reading, monitoring the effectiveness of their own performance and pattern synthesizer, i.e., processing of visual input and keeping it in the short term memory to build up a macrostructure (Rost, 2013).

A validation of the theoretical construct of reading, then, might involve having test takers report the processes, either introspectively or retrospectively, they use while responding to test tasks. If the tasks do reflect the discourse processing set out in the theoretical model, then we can claim our inferences about the test scores to be valid.

2.2.2 Before the test events: Context validity. Context validity, in Weir's (2005a) words,

> … is concerned with the extent to which the choice of tasks in a test is representative of the larger universe of tasks which the test is assumed to be a sample. This coverage relates to linguistic and interlocutor demands made by the task(s) as well as the conditions under which the task is performed arising from both the task itself and its administrative setting" (p.19).

The definition given here is somewhat parallel to what others have called content validity. Farhady (2012) for example, claims that content validity "refers to the correspondence between the content of the test and the content of the materials to be tested" (p.38). As it would not be possible to include the whole content, a representative sample of the content should be included in the test. Weir's definition of context differs from the traditional understanding of content by the inclusion of the social dimension of language use. In his view, the social setting in which the test taker is expected to use the language delineates the range of communicative tasks that need to be replicated in the test. This *situational authenticity* can be achieved by investigating the criterial features of the target language use domain closely and including them as much as possible in testing (Douglas, 2000; Weir, 2005a).

Task settings and task input and output are the two subheadings under context validity. Task setting describes the parameters of the task:

- response method: the type of answer expected (selected response or constructed response)
- weighting: the points assigned to test items (some items may be weighted differently depending on the processing demands on the test takers)
- knowledge of criteria: information on criteria that affects scoring (for example, spelling or punctuation mistakes may be penalized)
- order of items: depending on the type of reading activated –careful global, expeditious global, etc.— the order of the items may follow the order of the text or not. In careful global reading, for example, reading is seen as a cumulative process as the information being read adds up to the meaning constructed so far. In this type of reading it is advised to set the questions in the order of the text (Khalifa & Weir, 2009)
- channel of presentation: decision on whether non-verbal information will be included in the test
- text length: decision about the length of the reading text
- time constraints: decision about the time given to read each text.

The second set of parameters are related to the linguistic demands of the test tasks. Khalifa and Weir posit that the linguistic demands in a test need to be as similar as possible to those made by equivalent tasks Industry life language use at the level of performance which is being targeted if generalizations are to be made from test performance to language use in the future domain of interest (p. 104). The linguistic demands are specified as

- overall text purpose
- writer-reader relationship
- discourse mode

- functional resources
- grammatical resources
- lexical resources
- nature of information, and
- content knowledge.

2.2.3 After the test events: Scoring validity. Weir (2005a) considers reliability as one form of validity. It is a quality that is derived from the scores; hence, it is called scoring validity in this framework. He defines it as "the extent to which test results are *stable over time*, *consistent in terms of the content sampling* and *free from bias*" (p.23) (emphasis original). Several categories of reliability are identified in the literature such as test-retest reliability, parallel forms reliability, internal consistency and marker reliability (APA, 1999). Test-retest reliability is obtained by analyzing the scores obtained from the two administrations of the same test to the same test taker population. The scores give a correlation coefficient between -1 and +1, indicating lack of reliability on the former and perfect reliability on the latter. Several reservations were made on test-retest reliability method Anastasi, 1988) and parallel-forms reliability is preferred over it (Weir, 2005a). In parallel forms reliability, alternate forms of a test are given to the same test taker population and the results of the two tests are compared statistically to achieve a correlation, the square of which gives an estimate of the degree of overlap between the two test forms.

Internal consistency measures include methods such as split-halt reliability which is a statistical comparison of the test taker's scores on one half of the items with the other half of the items. The correlation of the two scores gives a reliability estimate. In cases where the items in one half are not equivalent to the other, other methods of split-half correlations are calculated such as KR20 or Cronbach's Alpha [If these are important enough to mention in your manuscript give some details related to them.].

Marker reliability in tests of speaking or writing is another measure that is carried out through statistical analysis of the ratings. Consistency in marking is sought in two ways: (1) intra-rater and (2) inter-rater reliability. Intra-rater reliability is the consistency of the marker within herself; that is, whether or not the marker who is confronted with the same quality of performance in two or more instances, gives the same marks. The inter-rater reliability refers to the consistency of the marks given to the same quality of performance by different markers. In both cases, correlation analysis is carried out to rate consistency.

The last parameter in scoring validity is grading and awarding, which refers to setting cut-off scores for expected proficiency levels (grading) and re-examination of borderline performances in order to make sure that the results are fair.

2.2.4 Criterion-related validity. Criterion related validity is the extent to which the test scores correlate with a suitable external criterion of performance. There are two types of criterion-related validity: concurrent validity and predictive validity. Bachman (1990) explains concurrent validity as the correlation of the test scores with another measure of performance (criterion) taken at the same time. It can also be teachers' evaluation or self-evaluation of the students; however, with these type of evaluations correlation may be low (Alderson et al., 1995; Weir, 1983).

Predictive validity is about the predictive power of an instrument in revealing test taker's future performance on a job or academic subject. It is somehow problematic to establish the predictive validity of a test since correlating test scores with later performance proves to be difficult due to confounding variables (Banerjee, 2003).

2.2.5 Consequential validity. The three parameters mentioned under consequential validity are:

- Impact on institutions and society
- Washback on individuals
- Avoidance of test bias.

On impact of the test on institutions and society, Messick (1989b) claimed that the potential and actual social consequences of test interpretation and use should support the intended purpose of using the test and that they should be consistent with social values. He said,

> [b]ecause the values served in the intended and unintended outcomes of test interpretations and test use both derive from and contribute to the meaning of test scores, the appraisal of social consequences of testing is also seen to be subsumes as an aspect of construct validity" (p. 18).

Washback of the test is another perspective to consider as a social consequence of testing. Weir quotes Hamp-Lyons (1997, in Weir, 2005a) on washback, who claims that the tests affect not only the test taker but the society and the education system therefore the tests need to be evaluated from the stakeholders', that is, learners', teachers', parents', government and official bodies' and the marketplace's, point of views as well.

Bachman (1990) identifies test bias as the differences between subgroups of test takers which may affect test performance, and consequently, undermine validity claims of a test. He claims that if systematic differences between test scores of subgroups are due to some individual characteristics but not the ability that is tested, then there is test bias. He lists four sources of test bias: cultural background, background knowledge, cognitive characteristics, and native language, ethnicity, sex and age. Weir (2005a) claims that avoidance of test bias is possible through carefully set guidelines of item writing and test development processes.

2.3 Other Approaches to Validity Inquiry

One conceptual study on validation is by Haertel (1985), who focused on the construct validity of criterion-referenced tests. He argued that *multifaceted inquiry* is needed to generate evidence for the meaningfulness of the instrument in making interpretations about the performances. One exception, he claimed, was when tests are used for summative purposes, i.e. to rank examinees, in which case, correlation studies could suffice to establish validity.

According to Haertel (1985), in criterion-referenced testing construct validation is required to make meaningful interpretations from test scores. The researcher emphasizes the need to gather different kinds of evidence – both theoretical and empirical – using a framework. Accordingly, the first step would be to define the instructional outcomes that are intended to be tested as achievement constructs in psychological and behavioral terms. This means, as well as the knowledge and skills that the construct necessitates, the definition should include how it is related to other constructs in the curriculum. After establishing the intended outcomes through construct definition, a smaller domain of testable outcomes was defined. Afterwards, a sampling process from the list of testable outcomes was carried out to assemble the test, guided by a reconciliation between practical limitations and objectivity concerns. Haertel (1985) encourages the use of empirical analyses such as regression. He claims these studies would help to validate the test as well as provide a sample of behavior representative of different levels of proficiency. To conclude, the researcher claims that such a validation study would not only yield better, reliable tests, but also help develop assessment batteries that are congruent with the needs of educational research.

Chapelle (1998) provides her rationale on validation in a way similar to Messick's (1989b) conceptualization of validity:

> Sufficient justification of the interpretations made from test performance in an operational setting is needed so that tests can be used appropriately for decision making in educational contexts or for theory construction in research settings. The process of securing sufficient justification is validation (p.49).

She views the evolved conception of validity in the 1980s and 1990s similar to the interactionalist construct definition. According to the interactionalist perspective, the trait that is being assessed and the context are closely related, and a test taker's performance is influenced by the context in which it occurs. Linking this definition to Messick's definition of validity, she claims that justifications for the interpretation and use of a test need to be supported with empirical studies showing that test performance does actually reflect the intended construct. Following Messick's four-cell matrix, she also mentions how to provide evidence related to the relevance and utility of testing: this evidence would show how useful a test is in achieving objectives in a particular context (Chapelle, 1998). The researcher concludes her paper claiming that the current approach to validity inquiry will have reflections on second language acquisition and language teaching research. She emphasizes that assessment batteries used in instruction and research should be "subjected to the processes of validity inquiry to reveal the quality of any given operational setting for producing the relevant signs and samples of learners' performance" (p. 64).

Kane (2011) proposes a different approach to validation, which he calls the *argument-based approach* (1992). Although he acquiesces Messick's (1989b) view on the need for justification for score based decisions, he proposes a more practical framework for validation. Messick's validity definition integrates all validation models into one that is based on construct validity. Kane claims that this view calls for multiple theoretical perspectives and different types of evidence for the interpretation and use of the test scores, which is a burdensome requirement for a researcher (2013). Hence, he proposes validity inquiry to be carried out in two steps: "specification of the proposed interpretations and uses of the test scores as an interpretive argument, and the evaluation of the plausibility of the proposed interpretive argument" (2011, p. 3).

An interpretive argument is about the rationale in drawing conclusions and making decisions based on the scores from an assessment. In other words, an interpretive argument comprises inferences about the quality of a performance. If one accepts scores to be indicative of an expected performance level within a domain, then it would be possible to make generalizations from those scores within a context, according to Kane (2011).

The second step proposed by Kane, is the validity argument which "provides an evaluation of the interpretive arguments coherence and plausibility of its inferences and assumptions" (2013, p. 8). He argues that the interpretive argument can be used as a framework for validation by specifying the inferences and assumptions that need to be evaluated.

2.4 Studies Utilizing Frameworks/Models

Weir and Khalifa (2008b) used the reading model specified in Weir and Khalifa (2008a) to examine two adjacent proficiency level exams, PET (B1 level of CEFR) and FCE (B2 level of CEFR), of the Main Suite General English examinations.

The reading model the researchers used, and which is also used in this present study to define the reading concept, is called the cognitive processing reading model (Figure 3) that consists of three main components: Goal Setter (left column), the Central Processing Core (middle section) and Knowledge Sources (right column). The choice of reading activity in the goal setter determines which processes will be prevailing in the central processing core, and which knowledge sources will be activated for comprehension. The goal setter specifies the purpose of reading as careful reading or expeditious reading.

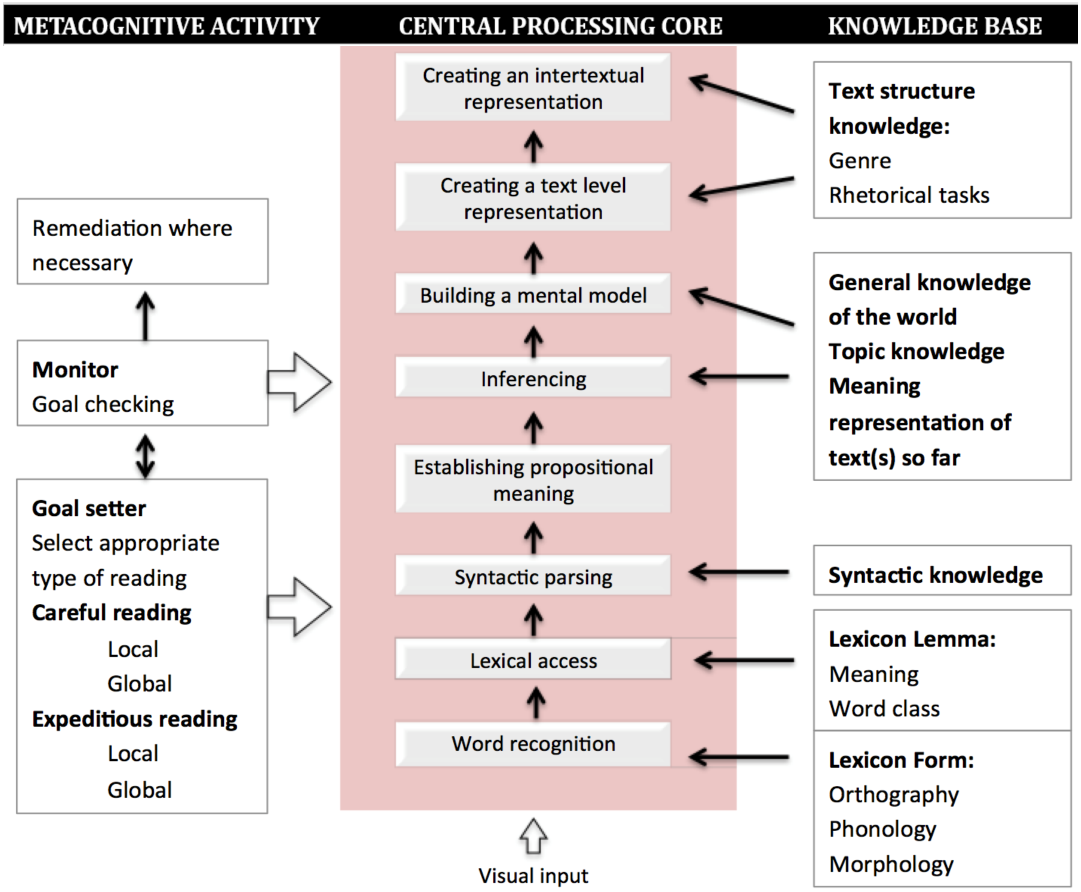


Figure 3 A model of cognitive processing in reading adapted from Khalifa and Weir (2009).

The model consists of three parts: **Metacognitive Activity** that defines the type of activities that the reader carries out, **Central Processing Core** that includes elements initiated by the activities carried out in the Metacognitive Activity, and **Knowledge Base** that refers to the types of knowledge that the reader brings into the reading process.

The type of reading that a reader decides to use are defined in the Goal Setter under Metacognitive Activity. Urquhart and Weir (1998) define reading at two levels: Careful Reading and Expeditious Reading.

Careful reading refers to comprehending the text completely. The reader may choose to do careful reading at the local or global level. Local careful reading is limited to understanding at word, clause or sentence level. It is generally used to resolve lexical ambiguity and identify pronominal references. The processes from *word recognition* to *establishing propositional meaning* in the central processing core are activated in local careful reading. In terms of knowledge sources, *knowledge of word forms*, *word meanings*, and *syntax* are activated in comprehension.

In global careful reading, on the other hand, the aim is to understand the main ideas by identifying the macro structure of the text, understand how ideas relate to each other and identify the writer's purpose (Weir & Khalifa, 2008a). In this type of reading, all the processes in the central processing core are activated as comprehension of both explicit and implicit information is necessary. In terms of knowledge sources, in addition to *knowledge of word forms*, *word meanings* and *syntax*, the reader uses his *knowledge of the world*, *topic knowledge,* and *knowledge of text structures* such as genre and rhetorical patterns.

Expeditious reading refers to quick and selective reading to access information in a text (Khalifa and Weir, 2009). Expeditious reading strategies involve scanning, skimming and search reading.

a) Scanning refers to reading selectively to identify specific words; hence, it is local expeditious reading. Here, the central processing core is activated from *word recognition* to *syntactic parsing* for singular words and *establishing propositional meaning* for word chunks or clauses.

b) Skimming, on the other hand, refers to going through a text quickly (Alderson, 2000a) at the global level, which means the reader will quickly and selectively create a *text level representation* if there is only one text, or create an *intertextual representation* is there is more than one.

c) Search reading is an activity that can take place at either local or global level. In search reading, the reader searches for a pre-determined topic. If the topic is found within a sentence, it becomes local reading. If it is found across sentences, it is global reading. Depending on the type of reading, related cognitive processes and knowledge sources will be activated as explained above.

The operations within the Central Processing Core, and Knowledge Sources as well as the interaction between the three parts of the diagram in Figure 3 are explained in detail in Chapter 4, where the conceptual and operational definitions of the reading construct are made.

Weir and Khalifa (2008b) used this reading model (Figure 3) to examine PET and FCE exams of the Main Suite General English examinations. The emphasis of the study was on

- the variety and complexity of the *reading types* demanded at B1/B2 levels
- the comprehensiveness of the *cognitive processes* covered by these two levels
- the cognitive demands imposed by relative *text complexity* in PET and FCE
- whether the cognitive processes elicited by these two exams resemble those of the reader in a non-test context (p. 11).

In terms of the variety of reading types, their investigation of the PET exam revealed that the test takers have to use both expeditious and careful reading skills at the local and global level. In FCE the tasks primarily focus on careful reading at the global level. They suggest that expeditious reading is encountered as search reading rather than scanning. For the levels of processing at B1 and B2 levels, both PET and FCE exams cover word recognition, lexical access, parsing, establishing propositional meaning, inferencing, and building a mental model[2].

Another area of the investigation of the study was on text complexity. The researchers posit that text complexity has an effect on the cognitive demands required of the reader. Indicators of text complexity could be whether the texts include high frequency or low frequency words, or whether it is short or long. Examining PET level texts revealed that they mainly consist of vocabulary that are familiar and simple. However, the FCE level texts include a broad range of vocabulary and more complex sentence structures, and content. Finally, the researchers claimed that these two exams elicit processes from the test takers that correspond to cognitive processes involved in reading in real life.

Krishnan (2011) set out to investigate the item types in the IELTS reading tests based on the model reading suggested by Khalifa and Weir (2009), which would provide validity evidence for the test. In particular, they examined the skills and strategies that the test takers employed to respond to 14 IELTS reading tests and whether the reading tests were adequate in testing reading ability comprehensively. The researcher emphasizes that it is important to identify what skills and strategies are you involved in the reading process so and to design valid instruments to assess the reading skill.

The study involved collection of both quantitative and qualitative data from two test takers who completed the IELTS reading tests under test-like conditions and while doing so they noted down the strategies that they employed in finding the answers. Analysis of the data revealed that the majority of the tests (77%) focused on careful reading as opposed to expeditious reading. However, as there were no time constraints on the test takers, it was not clear whether expeditious reading strategies could be employed dear and the test. The researcher suggests that careful reading items and expeditious reading items need to be tested separately to enhance the validity of the test. As a result, Urquhart and Weir's matrix was found to be meaningful in testing

[2] This list is part of the cognitive processes that constitute the reading model by Khalifa and Weir (2008a). This model is described in detail in Chapter 4.

reading ability comprehensively. However, the IELTS tests lacked items that required higher level cognitive processing: the majority of the items tested reading comprehension at the local level, which suggests that the tests may not be reflecting the actual reading ability of the test takers. The researcher suggests improving the test instrument by including a balanced number of items that's require both careful and expeditious reading at the local and global level.

A study that was conducted at the University of Minnesota (Chalhoub-Deville, Alcaya, & Mccollum Lozier, 2013) sought to define an operational framework to assess the reading proficiency levels of students in three languages and at three levels of proficiency. The researchers postulated that the theoretical models existing in the literature had "a global, all-encompassing perspective" (p.2), which seemed inapplicable to address particular needs in their specific testing situation. Hence, they started by defining language ability in their own testing context, and then, narrowed their focus to reading ability. Finally, they reviewed L2 reading research in order to adapt a model that would specify how the reading ability would be measured, how texts would be selected, what item types would be used, what the scores would mean and how those scores would be used (Chalhoub-Deville et al., 2013).

After reviewing the literature on language ability, particularly reading ability, and reading models, the researchers merged different perspectives to identify the elements of their own assessment framework. They described two major elements: *a text selection model* and *task criteria*. The *text selection model* consisted of four dimensions that were believed to be essential in defining text difficulty; namely,

- text types (wide availability – limited availability)
- the content (topics, cultural distance)
- the organizational characteristics (structural and rhetorical complexity), and
- the pragmatic features lexicon, function, sociolinguistic factors) (Chalhoub-Deville et al., 2013, p. 17).

The study concludes that the suggested framework would be utilized in text selection and in constructing text items, and that the *a priori* elements of the framework and their features would be continuously evaluated against the performance of the test takers.

Another study based on Weir and Khalifa's (2008a) model of reading was carried out by Katalayi and Sivasubramaniam (2013). In this study, they investigated the validity of a reading test with 50 multiple-choice items from the English state examination which was administered to Grade 12 students. They used a questionnaire adopted from Weir and Khalifa (2008a) to elicit the strategies and skills that the participants reported to have used during test taking. The questionnaire required the participants to choose from a list of strategies the one they have used to answer each question on the reading test. The questionnaire included the reading types as described in the aforementioned reading model: careful reading at global level, careful reading at local level, expeditious reading at global level, and expeditious reading at local level.

The researchers found that the emphasis of the test questions was on careful reading (64%) when compared to expeditious readings (36%). Moreover, more than half of the items targeted reading at global level than local level. They also posited that even though some test items originally targeted information at sentence level reading at global level was used as a general strategy. One question that could be raised about the methodology of the study is that it might be presumptuous to expect Grade12 students to be able to distinguish between reading strategies by looking at the test items and while trying to find a response to them.

2.5 Studies Based on Weir's Sociocognitive Framework

There are a number of studies in the literature that has used Weir's (2005a) framework as a basis for test validation. Within the framework of the new Test of English as a Foreign Language (TOEFL), Rosenfeld, Leung and Oltman (2001) carried out a study that investigated the academic tasks with regard to reading, listening, speaking and writing, that are important for achievement in academia. The researchers call it a *job analysis* which is needed to demonstrate the content validity of the assessment battery. The specific aims of the study were related to the validation efforts of the new TOEFL 2000 project framework, and included research questions that mainly fall in the categories of specifying the tasks that faculty members, graduate students and undergraduate students found important in satisfactorily completing the undergraduate and graduate level studies.

The task statements that were rated as *most important* by faculty members in terms of reading were "Reading text material with sufficient care and comprehension to remember major ideas" which received the highest rating, followed by "Read and understand written instructions/directions concerning classroom assignments and/or examinations" and "Read text material with sufficient care and comprehension to remember major ideas and answer written questions later when the text is no longer present" (Rosenfeld et al., 2001, p. 18). The first and third statements were categorized as a *learning* task, whereas the one in the middle was categorized as a *basic comprehension* task. In a similar manner, the graduate faculty respondents also rated the same three statements as the *most important* for competence in graduate studies with similar or higher mean scores.

The responses from undergraduate and graduate students were slightly different from those of the faculty members: according to undergraduate students, the most important task statements in the survey were, "Determine the basic theme (main idea) of a passage", "Read and understand written instructions/directions concerning classroom assignments and/or examinations" and "Read text material with sufficient care and comprehension to remember major ideas" (Rosenfeld et al., 2001, p. 32). The first two of the task statements mentioned were in the category of *basic comprehension*, and only the last task statement belonged to the *learning* category.

Finally, the reading task statements that were perceived as most important by the graduate student respondents were similar to those of the undergraduate student respondents. The graduate student respondents stated that, "Determine the basic theme (main idea) of a passage" was the most important task statement followed by "Read text material with sufficient care and comprehension to remember major ideas" and "Read text material and outline important ideas and concepts" (Rosenfeld et al., 2001, p. 36). The first task statement was in the category of *basic comprehension* whereas the last two were in *learning* category.

The researchers developed a pool of task statements observing the following criteria: that the statements were rated at least 4.0/5.0 by undergraduate or graduate faculty respondents, they rated 3.5/5.0 by undergraduate and graduate student respondents and that they have a mean importance rating that is in the top levels by either faculty or student respondents. The task statements that met the aforementioned criteria are summarized in Table 2.

Table 2 Reading task statements that meet criteria for inclusion in TOEFL 2000

Reading	Example Task Statement
Basic Comprehension	▪ Determine the basic theme (main idea) of a passage ▪ Read and understand written instructions/directions concerning classroom assignments and/or examinations
Learning	▪ Read text material with sufficient care and comprehension to remember major ideas and answer written questions later when the text is no longer present ▪ Read text material with sufficient care and comprehension to remember major ideas
Integration	▪ Compare and contrast ideas in a single text and/or across texts ▪ Synthesize ideas in a single text and/or across texts

The aforementioned statements were reported to be the most important and relevant for achievement in academic studies. They were also believed to be useful descriptors in curriculum design to guide nonnative speakers in improving their English language skills related to academic studies.

Another study on reading skills was conducted by Hudson (1996) relevant to the TOEFL 2000 project. The focus of the study was on academic reading from a communicative proficiency perspective in large scale assessment. Hudson explains that the context of discourse is a major underlying factor in understanding competence within communicative competence perspective. Accordingly, he asserts that the "… candidate should be allowed to demonstrate the ability to apply reading skills to a task in purposeful sociocultural context" (Hudson, 1996, p. 3).

Given the importance of context in the assessment of reading ability and the common agreement in current literature on the interactive nature of reading, Hudson (1996) makes a list of implications for reading assessment relating them to the four components of Messick's (1989b) validity definition; namely, construct validity, value implications, relevance/utility and social consequences (see Section 2.1 for a detailed introduction on the concept of validity). The first implication relates to the response formats. Hudson (1996) asserts that the response formats should expand beyond the multiple-choice format arguing that real life situations are far more complex than having to choose from a number of options; thus, constructed-response items are needed especially when the importance of context and purpose are taken into account in academic reading. This is closely related to construct validity.

A second concern over selected-response format is related to the unsubstantiated view of reading as comprising discrete subskills which can be isolated from the contexts they are applied. Hudson (1996) argues that skills overlap and they are applied differentially depending on the reading purpose. Hence, he argues that selected-response items by nature do not support the inter-related nature of reading context and reading purpose. Notwithstanding these arguments against the multiple-choice response format, Hudson (1996) suggests using a combination of selected- and constructed response items, which, he argues, may help overcome shortcomings of individual measurement instruments.

Another implication of the study is about creating authentic tasks in reading. Hudson (1996) points out that in academic settings reading is a skill that is complemented with other academic tasks such as writing in exams, or taking notes. Hence, a task-based approach to reading is advocated for reasons related to authenticity of the test task. Hudson (1996) also makes a reference to Messick's (1995) validity matrix claiming that task-based approach emphasizes language use in academic context and therefore this type of tasks will have positive value implications and social consequences.

Another study on the cognitive processes that underlie the academic reading construct was carried out by Weir, Hawkey, Green and Devi (2009). The purpose of the study was to clarify the link between the construct of reading and academic reading as practiced by students in a UK university. The specific aim of the study was to validate the reading component of the IELTS exam by examining the cognitive processes employed through participant retrospection. The study used Khalifa and Weir's (2009) model which "..accounts for the interactions between reader purpose, cognitive processes and knowledge stored in long-term memory" (Weir, Hawkey, Green, & Devi, 2009, p. 160). A retrospection form was designed for the test takers to complete immediately after responding to the questions on the IELTS Reading test. The questionnaire sought to clarify the sequence of reading activities, strategies for responding and information base for the response in three sections. In section 1, whether the participants read the text, and whether they used careful or expeditious reading strategies before reading the items were investigated. In section 2, the processes that the participants employed in responding the items were explored. These processes included matching words in questions with those in the text, using knowledge of grammar and vocabulary, trying to understand the meaning of a difficult word, etc. In section 3, the participants were asked to indicate where they felt they found the necessary information to respond each question; within a single sentence, by putting together information across sentences, by understanding how information in the whole text fits together, without reading the text or alternatively, whether they could not answer the question.

The results indicated that, previewing the text was a common strategy among the participants although the participants who scored higher reported less frequent use of this strategy than those who received lower scores. The participants used expeditious reading strategies consistently in all questions types. The type of reading in response to IELTS test items was found to be quite parallel to the approach specified in academic reading as reported by the students in a previous study (Weir, Hawkey, Green, & Unaldi, 2009). The response strategies that were reported to have been most frequently used were "quickly match words that appeared in the question with similar or related words in the text", "read the text or part of it slowly and carefully" and "read relevant parts of the text again". Finally, the third set of responses that elucidated how the participants found the information necessary to answer the questions, revealed that participants responded the tasks most frequently by "putting information together across sentences". Weir et al (2009) conclude that expeditious reading strategies are commonly used in the effort in answering test questions. The most commonly used strategy was matching words used in the question with similar or related words in the text. The study concludes that responding the IELTS test items, test takers approach to reading is consistent with academic reading that had been defined in the literature (Weir, Hawkey, Green, & Unaldi, 2009).

In this chapter, I provided the chronological developmental scheme of the concept of validity, a detailed account of the socio-cognitive framework employed in this study, and other studies based on validation frameworks and models. Reading ability and the test construct (of reading) which are the foci of this study are discussed in detail in Chapter IV, as part of the *a priori* validation of the reading test. In the next chapter, I present the research design, data collection instruments, and methods employed in data collection and data analysis.

CHAPTER 3

METHODOLOGY

The purpose of this chapter is to present the methods used in investigating the research questions of this study. The research design, participants, data collection and analysis procedures are discussed in detail.

The research questions addressed in this study are different in nature: the first research question is descriptive (Tully, 2014), i.e. it sought to describe reading ability by referring to its historical development and explicating current view of reading in relation to the local context. The second and third research questions are exploratory: the former investigated reader's cognitive processes through verbal protocols, and the latter investigated whether test scores are based on an appropriate criteria (Khalifa & Weir, 2009). Hence, as different methodological approaches were needed to answer each research question, they are presented separately in this chapter.

This study presents the validation procedures for a reading test, which involves investigation of contextual, cognitive and scoring validities of the test. The framework used in this investigation is derived from an evidence based validation model, namely, the socio-cognitive assessment and validation framework (Weir, 2005a). In this socio-cognitive approach to validation, the procedures carried out to investigate facets of the test are called *a priori* validation procedures, i.e. before the test events. In addition to these, there are *a posteriori* validation procedures that are carried out through the analysis of items after administering a test. In this study, the research questions investigate both *a priori* and *a posteriori* procedures: research questions 1 and 2 are related to the former and research question 3 is related to the latter aspect of validation. They are formulated as follows:

1) How is academic reading ability conceptualized and operationalized as a test construct?
2) What are the cognitive processes that underlie the construct of the reading test in retrospection and introspection?
3) To what extent do item parameters contribute to the validity claims of the test?

As mentioned at the beginning of this chapter, the research questions call for different methodologies. The first research question requires conceptual inquiry whereas the second and third research questions have empirical orientation. The methods used to answer each research question are explicated individually in the coming pages.

3.1 Research Design

This study has a mixed method research design as necessitated by the nature of the questions under investigation. Test development and validation processes demand both conceptual and operational definitions, and analysis of quantitative and qualitative data (Teddlie & Tashakkori, 2009). Specifically, this study is a complementarity mixed method study. The qualitative and quantitative methods were used to investigate overlapping aspects of the same phenomena (Greene, Caracelli, & Graham, 1989); namely, the cognitive processes used in answering a reading tests.

An assessment framework and a reading model guided the test development process in this study: The socio-cognitive assessment framework developed by Weir (2005a) and the reading model (Khalifa & Weir, 2009; Weir & Khalifa, 2008b, 2008a) which is an expanded version of an earlier study by Urquhart & Weir (1998). Weir's framework was used as a theoretical basis for both test development and validation in many contexts (Bannur, Abidin, & Jamil, 2015; Donaghue & Thompson, 2012; Nakatsuhara, 2011; Unaldi, 2010; Weir, Hawkey, Green, & Devi, 2009; Wu, 2011; Yanagawa, 2012). Within this framework, test development is defined in two main stages*: a priori* and *a posteriori* (Figure 4).

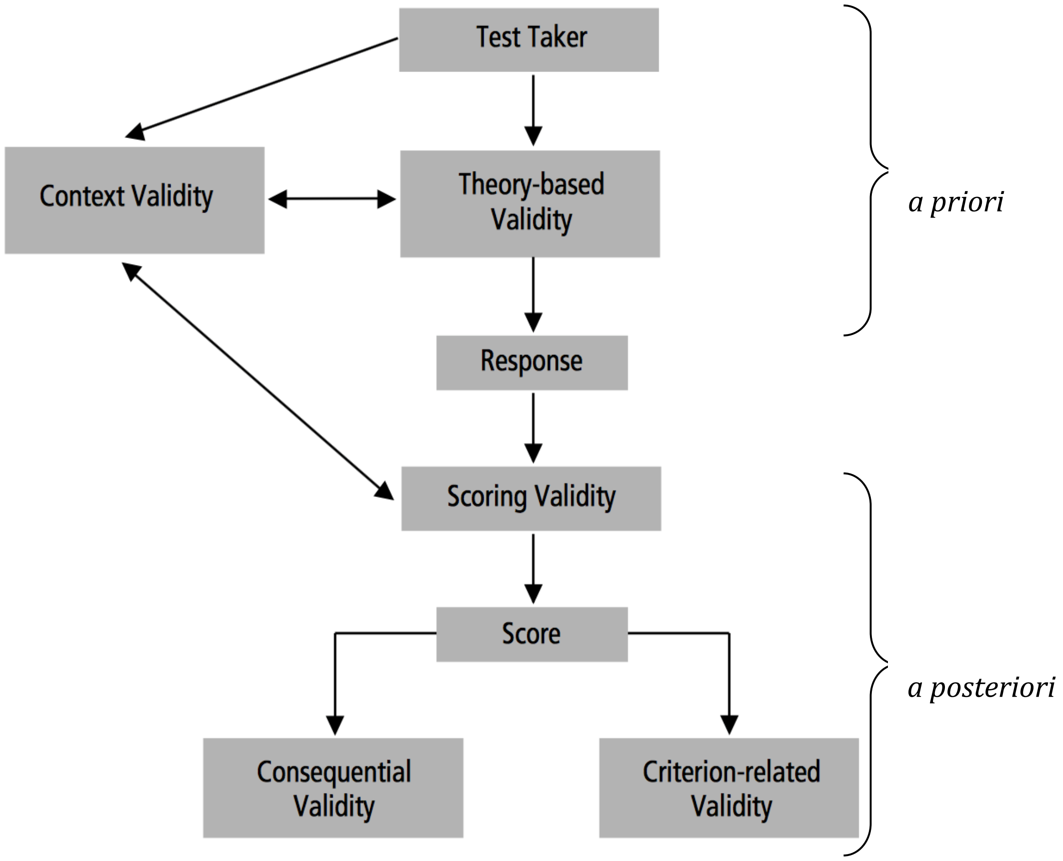


Figure 4 Framework for validating tests. Adopted from Weir (2005a).

In the *a priori* stage, first, conceptual definitions of the test elements and test environment are made (context and theory-based validity). Operationalization of the construct is completed through the specification of the test construct at this stage. Weir (2005a) maintains that

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

Hence, in *a priori* validation the focus is on generating a skill/language model consistent with the local needs. These may comprise, among others, analysis of communicative tasks used in teaching and testing, the expected linguistic range, and vocabulary range. Once these needs are established, current literature is reviewed to match them with a theoretical model in order to generate specifications for testing.

The second stage, *a posteriori*, is mainly concerned with generating empirical evidence on aspects of scoring after the test is administered (scoring validity). At this stage links are made between all elements of the model to make sure that the scores given are fair and meaningful (O'Sullivan & Weir, 2011). The empirical examination of the psychometric properties of test scores such as reliability and internal consistency measures provides evidence on the scoring validity of the test.

There are two other validation areas in this stage: criterion-related and consequential validity. Criterion-related validity is obtained by demonstrating that there is a relationship between the scores of a test and an external criterion i.e. an external test that is believed to measure a similar ability (Weir, 2005a). Consequential validity, on the other hand, stems from Messick's (1989b) validity theory and is mainly concerned with demonstrating "whether the social consequences of test interpretation support the intended testing purpose(s) and are consistent with other social values" (O'Sullivan, 2002, p. 22).

These two *a posteriori* validation procedures fell outside the scope of this dissertation as this study is focused mainly on the design stage i.e. the conceptualization and operationalization of the test constructs and gathering validity evidence on the construct validity *per se*.

3.2 Research Scheme

During the course of this study, I worked with a committee of four people, whose job was to write items for the proficiency exam under the supervision of the then director of the SFL. I was the leading member of the Research and Development Unit at the SFL whose role was to help establish a sound theoretical basis for the new exam, to guide the committee in using scientific approaches in item development procedures, and to follow the test development procedures as indicated by the assessment framework. I was also responsible for the preparation of all the documentation of the exam. The items of the reading test were prepared by the committee members, and vetted by me, the committee, and the administrators of the SFL, and edited by the committee members, if need be.

Having been assigned the task of implementing the socio-cognitive framework in the development of the reading test, I worked closely with the testing committee. We met regularly starting in September 2015, and continued to do so till the end of the 2016-2017 academic year.

The initial aim of the meetings was to discuss the implications of the needs analysis study for the proficiency test. The needs analysis study (for details please refer to Section 4.2.2) was carried out in all five faculties at METU in the 2013-2014 academic year. In the light of the requirements of different faculties and programs, and relevant literature on test development, the new conceptual framework, the reading model, and our students' needs were discussed. Previous conceptual and empirical research guided in reaching an agreement on a viable working model. Hence, the needs analysis study results, the literature, samples from the external examinations and the reading theory, draft test specifications were developed. It was decided that, during the course of item development and meetings, test specifications could be modified. The progression of the study in accordance with the guiding principles of the assessment framework are given in Figure 5.

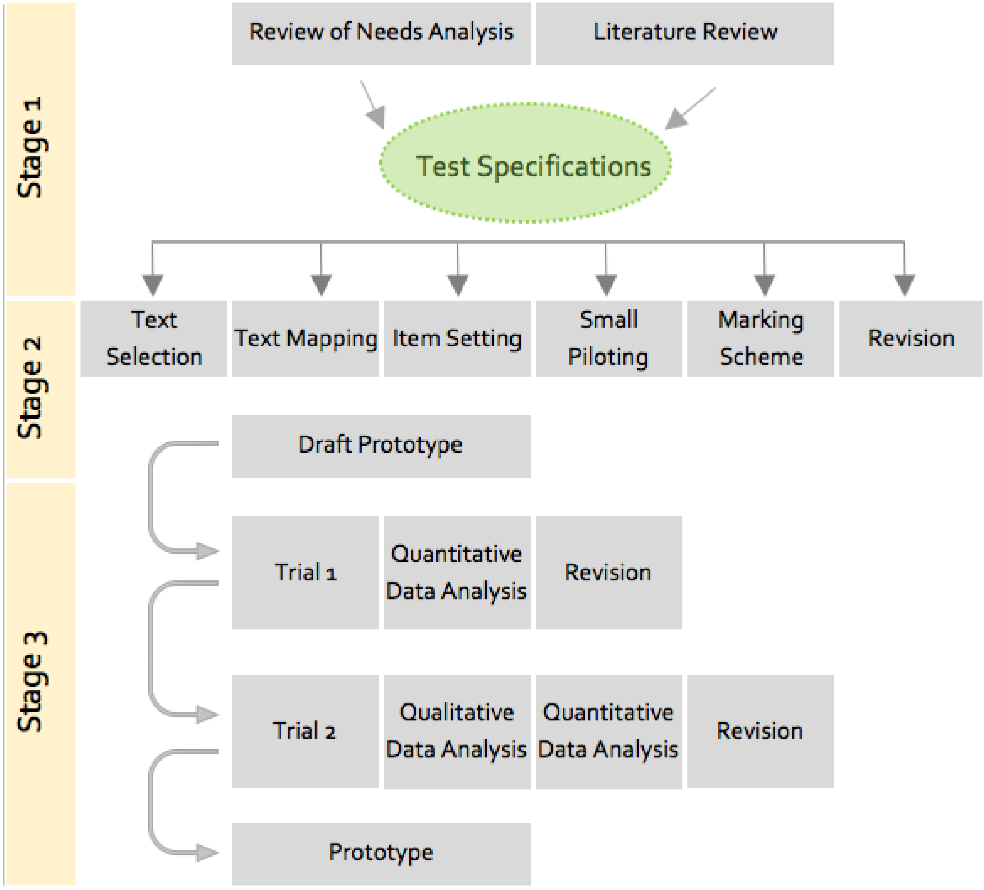


Figure 5 Progression of the study (adapted from Weir, Huizhong and Yan, 2000)

The committee members were responsible for producing test tasks in accordance with the test specifications. The development of the tasks was carried out through the course of text selection, item mapping, and item writing procedures.

Trial 1: After the final revisions, the first version of the reading test was compiled with four subtests (Text I + 8 items, Text II + 7 items, Text III + 8 items, Text IV + 7 items). For each reading text and its related items, a separate group of participants from the DBE were randomly chosen. The test was administered to the participants in their classrooms. The proctors were informed of the purpose of the administration of this test, and were given a short briefing, and written instructions (APPENDIX B). They were also provided with a report sheet (APPENDIX C) on which they were asked to record the period of time the students used to answer the test questions, and any questions they might receive from the participants during test administration.

The participants in each group were asked to fill in a retrospective protocol form as they answered the test items. I analyzed the scores of each task in accordance with the conventions of classical test theory (CTT). CTT, also called the true score theory, is the analysis of test items based on test scores. The statistics of CTT include measures of item difficulty, item discrimination, and test reliability, which are presented in Section 4.4. The results from the retrospective protocol form revealed the processes that the participants reported to have used while responding to the test. They also shed light on the skills and strategies that were put into use by the test-takers.

I shared all the results with the testing committee: we evaluated the results of the CTT analysis, i.e. whether item difficulty, reliability and discrimination values were within the expected value range (see Section 3.3.3). We also looked into the results of the retrospective protocol form analysis, and compared the reported processes with those that were proposed in the reading process model we have used in defining the test construct. Our discussions guided us in our decisions to keep items, to revise items for use in the future versions of the test, or to discard them.

Trial 2: In the period between April – June 2017, the second version of the reading test was administered to participants individually. At this stage of the study, the goal was to collect data to elucidate whether the cognitive processes employed during the test taking process reflected those in real life reading. .

The data was collected through verbal protocol procedures, and recorded. Then, it was transcribed and analyzed to reveal the type of cognitive processes used during the test (with relation to the different item types), and whether different item types call for the use of different strategies/cognitive processes. Moreover, test statistics were once again computed to reveal whether item difficulty, reliability and discrimination values were within the expected value range. After the administration of the second version of the test, items that were within the expected difficulty values, were then inspected for their reliability and discrimination power between test takers at different proficiency levels.

After transcribing all the recorded material, with the help of a software (MaxQda v.16) the coding of the transcriptions was completed. Once again, testing committee members and I got together to discuss, this time, the processes that the participants reported using and whether those processes were congruent with our expectations as test writers, and more importantly, with the reading process model, which specified the types of processes that would be activated for different reading purposes.

3.3 Research Questions

The methods used to answer each research question are given below.

3.3.1 Research Question 1. How is academic reading ability conceptualized and operationalized as a test construct?

This research question is associated with the context validity of the test. It has been argued that achieving context validity is problematic

> given the difficulty we have in characterizing language proficiency with sufficient precision to ensure the validity of the representative sample we include in our tests, and the further threats to validity arising out of any attempts to operationalize real-life behaviours in a test (Weir, 2005a, p.20).

However, some criteria were specified in approximating the real-life reading experience of university students. Accordingly, a reading model by Khalifa and Weir (2009) was used to define the reading tasks that were specified through the needs analysis process and supported with the findings in the literature (see Chapter 4 for the details of the reading model).

Having established the theoretical basis of the reading test, in other words, having conceptualized the reading skill as a test construct, test and item specifications were drawn in a joint effort with the committee. Test specifications inform the item writers how the reading construct will be operationalized in the test. In the case of this study, the reading construct is defined using Urquhart and Weir's (1998) taxonomy (Table 13) that accounts for the different purposes and processes of reading applicable in academic environments.

3.3.2 Research Question 2. The overarching research question 2 is: **What are the cognitive processes that underlie the construct of the reading test?**

This research question is related to the theory-based (or cognitive) validity, which eventually adds up to the overall –construct- validity of the test. The concepts used in the cognitive validity section of the framework (see Chapter 4 for details) refer to processes and resources that the test taker puts into use when taking a test. Those processes and resources are detailed in the reading model which is summarized above (see Section 3.3.1) and dealt with in more detail in Chapter 4.

In order to evaluate test takers' level of proficiency in a specific skill we need to break down that skill into parts that constitute it, which would allow the test developers to focus on these components in the tests (Weir, 2005a). In the present study, reading ability is investigated through the skills/strategies and knowledge sources that are claimed to constitute this ability according to the model proposed by Khalifa and Weir (2009)[3]. The researchers posit that "The cognitive validity of a reading task is a measure of how closely it elicits the cognitive processing involved in contexts beyond the test itself, i.e. in performing reading tasks in real life" (p. 3). The argument here is focused on establishing the generic cognitive processes that take place during reading, and sampling those processes in the reading tests. Weir et al (2009) suggest that it might be preferable for test/item writers to target test-takers' specific / underlying abilities if one can pinpoint exactly how particular item types target certain abilities. This is called the *subskills approach* which breaks down the reading activity into its components and the learner is expected to master each of these subskills to become a successful reader (Tracey & Morrow, 2012).

Alderson (2000a) takes a similar stance in applying the subskills in testing and posits that

> the validity of a test relates to the interpretation of the correct responses to items, so what matters is not what the test constructors believe an item to be

[3] The model is given in detail in Chapter 4.

> testing, but which responses are considered correct, and what processes underlie them (p.97).

Hence, for a better understanding of the processes that underlie the given responses to different types of test items, two protocols were used: the first one was carried out retrospectively, after responding the test questions, in written form and the second one was a verbal report of the processes as the participants responded the test questions.

3.3.2.1 Think aloud protocols. Verbal reports or think-aloud protocols have been widely used "as a method of identifying the mental processes that readers use to understand the printed word" (Anderson, 1991). In language testing, verbal reports are preferred instruments to understand the un-observable: test-takers' thinking processes as they set out to read a test prompt or answer a question. Green (1998) supports the view that verbal protocols are useful in gathering unequivocal evidence on the validity claims of assessment instruments.

There are different ways to elicit verbal reports: they can be retrospective and introspective: Retrospective method involves gathering the verbal reports immediately after the test-taker completes a given task whereas introspective reports are gathered during the test-taking process. Both methods were utilized in this study. The methods used in collecting retrospective and introspective data are explained in more detail under the related section of the research question. Hence, the second research question had two sub-questions:

a) What are the cognitive processes that underlie the construct of the reading test in retrospection?

b) What are the cognitive processes that underlie the construct of the reading test in introspection?

3.3.2.2 Research Question 2a. In order to answer research question 2a, retrospective data collection method was used. The instruments were Reading Test Version 1 (V1), and a retrospective protocol form. For the research question 2b, the Reading Test was revised (V2) and a think aloud protocol was used to collect data. The instruments and data collection methods for the two sub questions are presented in the next section.

3.3.2.2.1 Research instruments

3.3.2.2.1.1 Reading Test V1. The first version of the test had four subtests: each subtest consisted of a reading text, and either seven or eight items for each text. There was a total of 30 items in Reading Test V1. A number of different item types were included in the subtests: multiple-choice, multiple-matching, and true/false/not-given. Depending on the results of item statistics and analysis of the protocols, it was to be decided which item types should be kept in the final version of the test.

In terms of operationalization of the reading skill, careful reading at global level, expeditious reading at global level (skimming), careful reading at local level and expeditious reading at local level were prioritized. Search reading was not included in this version of the test.

3.3.2.2.1.1.1 The subtests. Each subtest contained a reading text and either seven or eight items. The texts were authentic texts taken from magazines, educational journals, proceedings etc. (e.g. Time Europe Magazine, BBC News). Written by native speakers of English for a general audience. However, some editing was done in order to ensure certain qualities related to text structure, lexical range and lexical characteristics. The topics of the texts were economy, personality traits, marine life and computer security.

After the texts were chosen in accordance with the specifications, they were analyzed for the difficulty levels and their vocabulary profile. The difficulty level of each text was examined using the readability function of a word processor (Microsoft Word, v.2016). The readability function includes two tests which are called the Flesch Reading Ease and the Flesch-Kincaid Grade Level. These tests provide information on the difficulty level of a text using core measures of word length and sentence length.. There is an inverse correlation between these two measures: As the score of the Flesch Reading Ease goes higher, the score on Flesch-Kincaid Grade Level goes down. On a 1 to 100 scale, 1 is the most difficult, and 100 is the easiest. Accordingly, the Flesch Reading Ease of the texts used in the four tasks were found to vary between 49.1 and 64.0.

Except for Subtest 2, the items related to the texts had a higher readability score (and, therefore, easier) than the texts themselves. In terms of the Flesch-Kincaid Grade Level, the texts and the items were found to be appropriate for grades seven to twelve (Table 3).

Table 3 Textual characteristics of the subtests

	Subtest I		**Subtest II**		**Subtest III**		**Subtest IV**	
	Text I	**Item s**	**Text II**	**Item s**	**Text III**	**Item s**	**Text IV**	**Items**
Number of words	652		979		662		1,127	
Number of paragraphs	6		10		8		9	
Average- sentences per paragraph	5.3		7.3		4.8		6.3	
Average- words per sentence	20.3		13.4		16.9		21.8	
Flesch Reading Ease	52.6	65.2	64.0	58.1	59.3	63.7	49.1	54.1
Flesch-Kincaid Grade Level	10.8	7.0	7.5	8.4	9.1	7.1	11.8	8.9

In terms of vocabulary coverage, an online program (http://www.lextutor.ca) was used to analyze the vocabulary profile of each text. A vocabulary profiler provides the frequencies of the words in a text and, therefore, makes it easier to understand whether a given text is appropriate for readers at a particular level of language proficiency. The corpus chosen for this analysis was the New General Service List (NGSL) (Browne, Culligan, & Phillips, 2013b) and The New Academic Word List (NAWL) (Browne, Culligan, & Phillips, 2013a) (for more information on the NGSL and NAWL, and the advantage for the present context over the other lists, see Section

4.2.2.5.3) . All the words in the four texts were analyzed using this corpus and the results are presented in Table 4. Accordingly, the first three lines of the table present the percentage of each text that is covered by the words from the first, second and third set of lists from the New General Service List (NGSL) and the percentage that is covered by the words in the New Academic Word List (NAWL).

Table 4 Vocabulary profile: Text coverage of NGSL 1, 2, 3 and NAWL (%)

	Text I	**Text II**	**Text III**	**Text IV**
NGSL 1	81.17	81.39	79.38	82.07
NGSL 2	88.64	86.75	86.61	88.29
NGSL3	90.88	89.17	88.76	92.13
NAWL	92.08	90.01	90.45	93.87

3.3.2.2.1.2 The items. Each of the reading subtest included either seven or eight items. Subtest 1 had multiple choice and matching items, Subtest 2 and 4 had only multiple choice items, Subtest 3 had yes/no/not given and multiple choice items. Each item type targeted certain skills and strategies that were indicated in the test specifications as part of the reading construct. The item types used in the tasks were **matching** (MAT), **multiple choice** (MCI), and **Yes/No/Not Given** (Y/N/NG).

Matching items consist of statements and associated responses. The test taker is expected to draw a correspondence between a statement and a response in accordance with the given directions. This type of item is used to test a number of reading skills such as understanding the main idea of a passage, understanding paraphrasing, and distinguishing between the main idea and supporting details (Haladyna & Rodriguez, 2004). The common use of a matching item in the present reading test is to ask the test taker to choose the most suitable headings that match the given paragraphs. A sample matching item from Subtest 1 is given in Figure 6.

Choose the most suitable headings for paragraphs A-F from the list of headings below (1-8). Write the appropriate number next to the blanks provided. BE CAREFUL, there are more headings than you need.

1. Paragraph A: __
2. Paragraph B: __
3. Paragraph C: __
4. Paragraph D: __
5. Paragraph E: __
6. Paragraph F: __

1. Why families of migrants want to make the best use of remittances
2. Reason for the difficulty to estimate the true figures of remittances
3. Why mass migration has increased so rapidly in the past few years
4. Efforts to make the most of remittances in the receiving countries
5. The undesired impacts of remittances on the receiving countries
6. What remittances are used for in the receiving countries
7. The motivation for migration from Morocco to France
8. The underlying motive for keeping a record of remittances

Figure 6 Matching item example

Multiple choice items consist of an incomplete statement or a question, and either three, four or five options. The test-taker is expected to choose the option that completes the statement or answers the question correctly. In the reading test a three-option multiple choice format was used as it was found to be optimal.

Rodrigues (2005) did a meta-analysis of research carried out in the last 80 years by reviewing both empirical research and narrative and theoretical reviews related to the optimal number of multiple choice options. He claims that the results bear implications for the validity arguments of the interpretations of test scores: using three-option multiple choice questions reinforces some aspects of validity arguments. The synthesis of past research proved that using more than three options does not improve the item much, and ends in implausible distractors, and thus, test takers continue using distractor deletion method, which makes his argument stronger.

He summarizes the practical arguments for the use of three-option multiple choice questions as follows:

- Less time is needed to prepare two plausible distractors than three or four distractors.
- More 3-option items can be administered per unit of time than 4- or 5-option items, potentially improving content coverage.
- The inclusion of additional high-quality items per unit of time should improve test score reliability, providing additional validity-related evidence regarding the consistency of scores and score meaningfulness and usability.
- More options result in exposing additional aspects of the domain to students, possibly increasing the provision of context clues to other questions (particularly if the additional distractors are plausible) (Rodrigues, 2005, p.11).

There are criticisms against the use of multiple choice type item formats in testing. It has been claimed that multiple choice items elicit lower level cognitive behavior (as opposed to constructed response items) or that test-takers may use some techniques irrelevant to reading ability to arrive at an answer. These downsides could contribute to construct irrelevant variance in test scores (Osterlind, 1998).

Advocates for the use of multiple-choice type items claim that the drawbacks associated with this item type can be overcome by adequate training of the item writers (Haladyna & Rodriguez, 2004), and that multiple choice items are versatile and can be used to test a broad range of reading skills (Green, 2014).

In the decision to use – primarily – multiple-choice items for the reading test a number of factors were considered: The committee had been using this type of items for many years, and they are experienced in creating high quality items of this sort; the stakeholders – test takers and instructors – are used to this format, and the majority of the practice materials for the test are in this format; there is very limited time for scoring the papers, which makes it almost impossible to clerically mark all the items;

and finally, due to the nature of the test, a high-stakes test with summative purposes, it is important for the administration to have the test objectively scored, as much as possible.

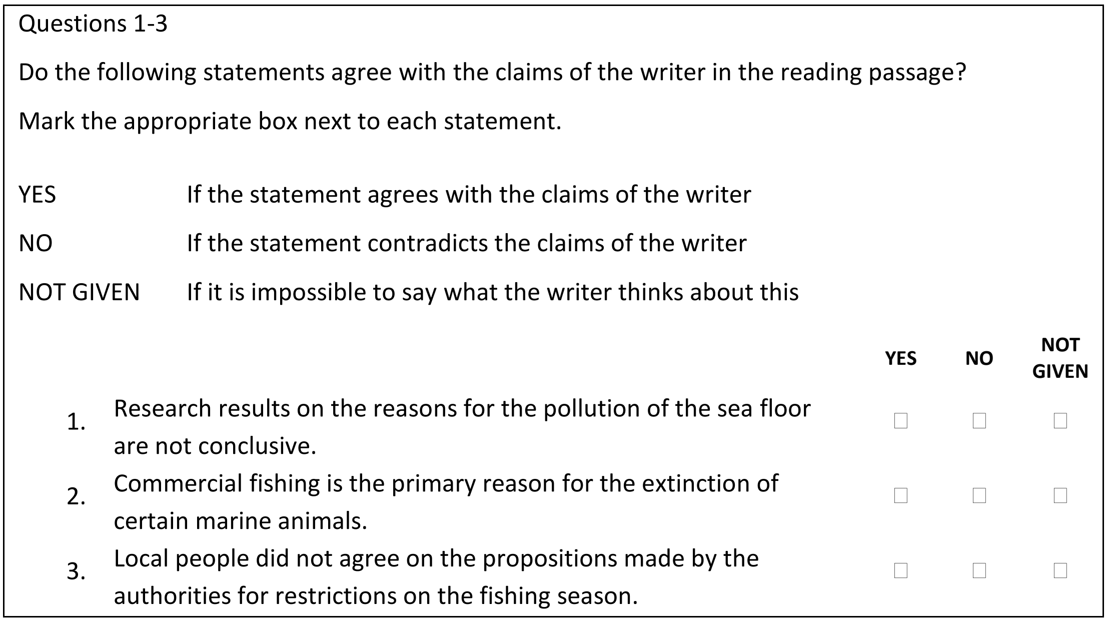
In the reading tests, multiple choice items are used to test skills and strategies such as understanding explicitly or implicitly stated ideas, guessing meaning of unknown words from context, understanding rhetorical strategies, organization of a text, and writer's attitude, purpose, or the communicative function of text (for examples see https://www.ets.org/Media/Tests/TOEFL/pdf/SampleQuestions.pdf, https://takeielts.britishcouncil.org/prepare-test/practice-tests/reading-practice-test-1-academic). An example of a multiple choice item from the Reading Test V1 is given below.

In paragraph G, the writer refers to some university students in Central Europe to imply that __________.

a) the number of hackers has reached alarming levels there
b) new training programs have already started in that part of Europe
c) training children about security in the virtual world is a challenging task

Figure 7 Multiple choice item example

Yes/No/Not Given item type includes a statement that the test taker decides whether it is true or false according to a given text. The third option "Not Given" is used to reduce the chance of guessing correctly. True/False/Not Given items can cover a wide range of content. It has been claimed that they can be used to test comprehension of both lower level reading skills such as understanding propositional meaning at sentence level and higher level skills such as generalizations, relationships between events, people, etc., and predictions (Osterlind, 1998). A sample item for *Yes/No/Not Given* is given below: (Figure 8).



Questions 1-3

Do the following statements agree with the claims of the writer in the reading passage?

Mark the appropriate box next to each statement.

YES	If the statement agrees with the claims of the writer
NO	If the statement contradicts the claims of the writer
NOT GIVEN	If it is impossible to say what the writer thinks about this

		YES	NO	NOT GIVEN
1.	Research results on the reasons for the pollution of the sea floor are not conclusive.	□	□	□
2.	Commercial fishing is the primary reason for the extinction of certain marine animals.	□	□	□
3.	Local people did not agree on the propositions made by the authorities for restrictions on the fishing season.	□	□	□

Figure 8 Yes/No/Not Given

3.3.2.2.1.2 Retrospective protocol form. A retrospective protocol form (see APPENDIX A) was used to collect data from the participants as they took the reading test. The form aimed to collect information on the strategies utilized during the test taking process. The form was taken from a study by Weir, Hawkey, Green and Devi (2009) in which cognitive processes underlying the academic reading construct in an IELTS test were investigated. The original form was in English. As the questionnaire was planned to be given to students at different proficiency levels, and comprehension of the questions was important in obtaining a reliable response, I translated the questionnaire into Turkish. Two experienced instructors in the same institution back-translated the form, individually. Their back-translations were compared with each other, and then with the original document. There were some minor differences between the back-translations: using synonymous English words when translating a word from Turkish and a small difference in one sentence structure. The Turkish form was finalized after inspecting the back-translations.

There were three sections in the form: The first section was about the sequence of reading activities. Here, information regarding the test takers' strategies *before* looking at the items were sought; that is, whether they read the text before looking at the items and whether they employed careful or expeditious reading strategies. The three choices in this section were:

1) read the text or part of it slowly and carefully (corresponds to careful reading)
2) read the text or part of it quickly and selectively to get a general idea of what it was about (corresponds to expeditious reading / skimming)
3) Did not read the text.

The second section sought information on the processes that the test takers were engaged in *while* answering each of the items. The test makers were allowed to choose more than one item as a number of knowledge base and cognitive strategies could be involved while locating the correct answer. There were eleven items listed in this section. They were:

1) match words that appeared in the question with exactly the same words in the text
2) quickly match words that appeared in the question with similar or related words in the text
3) look for parts of the text that the writer indicates to be important
4) read key parts of the text such as the introduction and conclusion
5) work out the meaning of a difficult word in the question
6) use my knowledge of vocabulary
7) use my knowledge of grammar
8) read the text or part of it slowly and carefully
9) read relevant parts of the text again
10) use my knowledge of how texts like this are organized
11) connect information from the text with knowledge I already have

The third section of the form consisted of items that sought to investigate where the test taker found the necessary information to answer the test questions. The options were:

1) within a single sentence
2) by putting information together across sentences
3) understanding how information in the whole text fits together
4) I knew the answer without reading the text
5) I could not answer the question

3.3.2.2.2 Participants of Reading Test V1. Participants included students studying English at the DBE in 2015-2016 academic year. At the DBE, students are grouped into four levels, beginner, elementary, intermediate and upper-intermediate, in accordance with the scores they receive from a placement exam. In addition, there is a repeat group which is made up of students who fail to pass the proficiency exam the previous year. And finally, in the 2015-2016 academic year, a new program was piloted in the beginner group which was called the pilot group. As a result, there were six level groups, and students from all groups were included in the study. At the time of the administration of Reading Test V1, the Beginner and Pilot group students had another 120 hours of instruction hours to complete before they were allowed to take the exam. However, they were included in this study as a choice of policy. The weekly instruction hours for each group were different.

- Beginner and Pilot group: 30 hours
- Elementary and Intermediate group: 20 hours
- Upper-Intermediate: 15 hours
- Repeat group: 15 hours

Students assigned to a level group at the beginning of the fall semester continued in a higher level group in the spring semester. There were 2918 students in 150 classes at six level groups at the DBE in the spring semester of 2015-2016 academic year. The groups and the distribution of students into these groups were as follows (Table 5):

Table 5 Student numbers in 2015-2016 academic year

Fall	Beginner	Elementary	Intermediate	Upper-Intermediate	Repeat	Pilot Beginner
Spring	Pre-Intermediate	Intermediate	Upper-Intermediate	Advanced	Repeat	Pilot Pre-Intermediate
	681	868	574	296	355	143

3.3.2.2.2.1 Sampling. The classes included in this study were randomly drawn from this student population: four classes from each level group, making a total of 24 classes and 400 students. The number of participants from each level group was almost equal (see Table 6) making the sampling method disproportionate random sampling.

There were four reading tasks to be distributed to 24 classes (four classes from each of the six level groups). Accordingly, each task was given to about 100 students in total, making up 400 participants for the total reading test. The distribution of the tasks within the level groups and number of test takers for each task is given in Table 6.

Table 6 Phase 1-Participant level groups

Group levels	**Subtest I**	**Subtest II**	**Subtest III**	**Subtest IV**	**Total**
Pre-Intermediate	16	15	18	18	67
Pilot Pre-Intermediate	20	18	18	17	73
Intermediate	18	16	19	18	71
Upper-Intermediate	13	17	16	16	62
Advanced	18	15	16	17	66
Repeat	16	16	18	11	61
Total	101	97	105	97	400

Majority of the 400 students who participated in this part of the study were Turkish (n=394). The other nationalities were Kurdish (n=3), Azerbaijani (n=1), Arab (n=1) and one unknown. There were 202 female, and 197 male students, and 1 no answer. The age mean was 19.7 (Table 7). As the proportion of foreign students in the study is negligible (around 1.5%), the discussion will be carried out with relation to Turkish students.

Table 7 Phase 1- Participant information

Mother tongue		**Gender**		**Age (Range: 18-21)**	
Turkish	394	Female	202	Mean	19.7
Kurdish/Zazaish	3	Male	197	Median	19.0
Arabic	1	No Answer	1	Mode	19.0
Azerbaijani	1			SD	1.23

3.3.2.2.3 Data collection and data analysis procedures for Reading Test V1. In order to investigate the cognitive processes underlying the reading construct, the reading test and the retrospective protocol form were administered to the participants on a designated day, in the spring semester of 2015-2016 academic year. Due to the dispersed nature of the campus settlement, and the need to administer the test and the form on the same day to 24 classes, 10 instructors (including me) participated in the data collection process. On the data collection day, the assigned instructors had a meeting with me where the procedures to be followed were presented and discussed with them. Each instructor was also given an instruction sheet (APPENDIX B) explaining how to carry out the procedure. As there were only two foreign students in the chosen classes, and beginner and elementary level groups were also included in the study, the administration of the test was carried out in Turkish. Hence, the instruction sheet was also prepared in Turkish. The two foreign students within the sample were given the English (original) version of the questionnaire. The instructors were also provided with a report sheet (APPENDIX C). They were asked to note down

(i) the questions asked by the students

(ii) the problems they encounter during the administration of the test,

(iii) the number of students taking the test,

(iv) the time needed to take the test and fill in the form.

On the designated hours, assigned instructors carried out the data collection procedure in the classes. Each test was given to one class from each level group, to an average of 100 participants per test, and 400 participants in total.

The participants were asked to fill in the retrospective protocol form right after they answered the questions on a reading test. For each reading question they answered, they also answered three different questions on the retrospective protocol form. The procedure lasted for one class hour (50 minutes), in some classes (especially in higher level groups) students finished earlier (the earliest in 15 minutes). Both the tests and the forms were collected by the assigned instructors, and returned to me.

For the analysis of the data, I marked the reading tests and then entered both the scores and the data from the retrospective protocol form into Microsoft Excel (v.2016). Later, I transferred the data into IBM SPSS Statistics (v.23). I analyzed the test scores using CTT methods in the summer of 2016. The data on the protocol form was entered into a Microsoft Excel (v.2016) file and descriptive statistics were computed for all parts of the retrospective protocol form. In order to understand whether level of proficiency had an effect on the use of skills and strategies while answering test questions, the results were analyzed once more separately for participants with having lower and higher level proficiency. Mann Whitney U tests were computed to reveal whether the differences were significant or not.

Beginning in September 2016, I and the testing committee got together to discuss the results (see Section 4.4 for the results of the CTT analysis). The CTT analysis of Subtest 3 revealed that the four out of six Yes/No/Not Given questions were too easy (M=74%). Moreover, the content of the reading text used with those questions added to the problem: there was too much factual information; therefore, the questions were testing the comprehension of explicit information only. As text difficulty and item difficulty are determinants of task difficulty (Grotjahn, 2001), it seemed plausible not to include this item type in the future exam versions, and to be more selective in choosing texts where there is a mixture of both concrete and abstract information, or some argumentation which would better yield to questions to test higher level skills.

The other subtest that was founded to be in effectual was Subtest 1. There were 8 items in Subtest 1, and 5 of them were below the expected facility value (.40). The overall mean value was, apparently, lower than 40%, it was 38%. The committee inspected the text and items again, and found the following problems: the content of the text was irrelevant for our test taker population (remittances sent home by legal/illegal migrants), and at least two of the headings were not worded appropriately.

The discussions ended with decisions regarding text quality (more detailed and careful selection of texts), item types (no yes/no/not given items), and making inquiries about how best to test expeditious reading, as the task that was intended to test it (Subtest 1), did not work well and was discarded.

3.3.2.3 Research Question 2b. What are the cognitive processes that underlie the construct of the reading test in introspection?

In the second stage of the examination of the cognitive processes, verbal analysis method was used to gather data from the participants. Information on the revised reading test, the participants and data collection procedure are given below.

3.3.2.3.1 Research instruments. The research instruments were Reading Test V2, and think aloud protocols.

3.3.2.3.1.1 Reading test V2. The second version of the reading test was administered to participants in the spring semester of 2016-2017 academic year. The test consisted of four tasks, each with a reading text, and a total of 30 items. Two of the tasks from Reading Test V1 were discarded and two new tasks were introduced, with new items and a new item type: open-ended item. The reason for introducing this new item type was due to the need to test expeditious reading skills more efficiently.

In the first version of the reading test, expeditious reading was tested through the matching items, and the committee expected the test takers to do skimming as an expeditious reading strategy. However, after examining the results of the retrospective protocol form, it was found that many test takers (close to 1/3) employed careful reading strategies instead. Hence, I and the committee reviewed the literature, and the testing practices of similar institutions (Language Preparatory Schools at Bogazici University, Sabanci University, Bilkent University, etc.) once more to decide on a more efficient task to test expeditious reading skills. Further group discussions revealed that instead of selected response, constructed response type items could better measure expeditious reading skills. In testing expeditious reading, finding the location of relevant information and reading carefully to extract the meaning of searched information were the main operations expected from test takers. It was decided that presenting the answer in a key option, together with two other distractors would not generate processing similar to that in real life, and a short answer format would be more suitable to test this skill efficiently.

Table 8 Textual characteristics

	Task 1		**Task 2**		**Task 3**		**Task 4**	
	Text I	**Items**	**Text II**	**Items**	**Text III**	**Items**	**Text IV**	**Items**
Number of words	715		1,127		979		2,067	
Number of paragraphs	6		9		10		33	
Average- sentences per paragraph	5.3		6.3		7.3		4.6	
Average- words per sentence	22.3		21.8		13.4		20.9	
Flesch Reading Ease	61.9	71.5	49.1	54.1	64.0	58.1	40.9	53.8
Flesch-Kincaid Grade Level	10.1	6.0	10	8.9	7.5	8.4	12.7	9.4

The new tasks had two new texts: Subtest 1 had a similar theme to the previous one, marine animals, and similar in length (715 words); Subtest 4 had a text much longer than the previous one (about 2 thousand words) in order to test the skill of expeditious reading more efficiently (the test taker will demonstrate her skill in using expeditious reading strategies by identifying the location of a predetermined topic in an extended text and read carefully to extract the necessary information). The text was about a seed preservation facility. The development of the new items in this new task were carried out in the same manner as specified in Stage 1 of the research scheme. The textual characteristics of the second version of the reading test are given in Table 8.

The vocabulary profile of each text was computed using similar procedures with Reading Test V1 (Table 9).

Table 9 Vocabulary profile: Text coverage of NGSL 1, 2, 3 and AWL (%)

	Text I	**Text II**	**Text III**	**Text IV**
NGSL 1	80.66	82.07	81.39	73.69
NGSL 2	88.50	88.29	86.75	86.41
NGSL3	92.57	92.13	89.17	92.05
NAWL	94.25	93.87	90.01	94.27

3.3.2.3.1.1.1 The items. This new compilation of the reading test had some differences from the first version in terms of item types. First of all, the committee decided not to use True/False/Not Given type of test items since the item statistics were not at expected level (see Section 4.3.1). Secondly, it was decided to introduce open-ended items to measure test-takers' ability in doing expeditious reading. Despite the challenges of scoring open-ended item type objectively and quickly, constructed response items have a certain advantage over the selected response item types Rauch and Hartig (2010) posit that constructed response type items replicate the teaching and learning processes better than selected response type items. In terms of the difficulty level, it has been claimed that selected response type items are generally easier than constructed response type items (Shohamy, 1984). The distribution of item types into texts is given in Table 10.

Table 10 Item types in the reading tasks

Text	Item	Response Type
1	1-6	Selected Response: Matching
1	7-8	Selected Response: Multiple-choice
2	9-15	Selected Response: Multiple-choice
3	16-22	Selected Response: Multiple-choice
4	23-30	Constructed Response: Short answer

3.3.2.3.1.2 Participants of Reading Test V2. The participants of the second phase of the study were students studying English at the DBE in 2016-2017 academic year. Those students who responded to the flyers posted on the bulletin boards were recruited for the study. There was a total of 27 students. Their level groups in the Spring semester were as follows (Table 11):

Table 11 Phase 2-Participant level groups

Group name	Student #
Intermediate	10
Upper-intermediate	8
Advanced	9
Total	27

The inclusion of intermediate level students was purposeful. The name of their level, intermediate, implies that they are not yet ready to take a proficiency exam that was aimed at the CEFR level of B2. However, they had only about 20 – 30 hours of instruction left before they were allowed to take the proficiency exam in June. Therefore, I believed, some of them might be at the borderline level, and therefore provide a good example of students at that level.

The pre-intermediate and pilot pre-intermediate students were not recruited as those students had another semester to complete (the extended term) before they were allowed to take the proficiency exam in August and therefore it had been assumed that they were not at the minimum expected level of proficiency to take the examination.

Demographic information about the participants in Phase 2 is given in Table 12.

Table 12 Phase 2-Participants' demographic information

Mother tongue	
Turkish	27

Gender	
Female	10
Male	17

Age (Range 18 – 20)	
Mean	19.1
Median	19.0
Mode	19.0
SD	0.39

3.3.2.3.1.3 Data collection and data analysis procedures for Reading Test V2. In order to investigate the cognitive processes underlying the reading construct as well as to reveal whether different item types call for the use of different cognitive processes and/or reading strategies, the second version of the reading test was produced between October 2016 and March 2017. This version of the reading test was administered to 27 participants individually in April, May and June 2017. The participants were recruited through banners posted on the bulletin boards of the language school. I gave each candidate general information about the nature of the study and an appointment was made. At most two appointments were given for one day, as each data collection session lasted for about two to two and a half hour. On the designated day, I met with each participant in a vacant classroom. I explained the overall aim of the study, informed students about the method of data collection and then carried out a short training session on how to think aloud while taking the test.

The training included an introduction of the procedure in writing, that is, answering a reading test and thinking aloud. First, I informed the candidate about the timing of the study. The training was about 20 minutes long and for taking the test, the candidates had two hours. Secondly, I set down with the candidate and handed him/her a copy of the instruction page (see APPENDIX D), which I read aloud and the participant listened to and followed the printed version at the same time. After reading instructions, I asked the participant to watch me while I demonstrated doing a think aloud with a sample question.

After my demonstration, I asked the participant to try thinking aloud with another sample question. Only after the participants felt confident enough to proceed, I handed out the actual reading test and started recording the procedure.

Each session was recorded using a Philips voice recording device. The files were later transferred to a computer to be transcribed. The sessions were transcribed and saved in a file using Microsoft Word (v.2016).

The analysis of the data was carried out in stages: The transcriptions were exported to MAXQDA (v.11) for qualitative analysis. Each participants' transcription was examined for explicitly verbalized strategies, or *moves,* as Cohen and Upton (2006) refer to them, that have a specific purpose and function. The coding scheme was based primarily on Pressley and Afflerbach's (1995) list, and Cohen and Upton's (2006) rubrics for coding reading strategies. In Pressley and Afflerbach's list, there were around 150 reading activities which were based on

> (a) planning and identifying strategies that help in constructing the meaning of the text,
>
> (b) monitoring strategies that serve to regulate comprehension and learning, and
>
> (c) evaluating strategies by which readers reflect or respond in some way to the text (Cohen and Upton, 2006, p. 4).

Cohen and Upton designed their own rubric also in three categories Reading, Test Management, and Test-Wiseness Strategies, containing 59 strategies. In the present study, the rubric from the Cohen and Upton study was used. In addition, some strategies which could not be fit into the existing categories, were added to the list. The final coding taxonomy list can be found in (APPENDIX G).

During coding, it was observed that the participants sometimes used more than one strategy while working on an item. In that case, all verbalizations related to an item were coded separately assuming that they were conscious references to communicative functions. One month after finishing the coding of all transcriptions, I randomly selected five transcriptions in their raw format and recoded them. There was

5 – 10% discrepancy with the first coding, which showed that the coding was carried out in a consistent manner.

Upon completion, I had a list of strategies, each allocated a frequency score, for each item type, i.e. vocabulary, critical reading and macro level comprehension, matching and search reading. I mapped those scores onto the relevant sections of the reading model to reveal the processes and the knowledge bases that were activated during test taking.

A second stage of analysis was the marking of the reading tests and analysis of scores according to CTT. The facility value, reliability, and discrimination power of each item was calculated and presented in Section 4.3.2.

3.3.3 Research Question 3. To what extent do item parameters contribute to the validity claims of the test?

This research question is associated with the scoring validity of the test. In testing receptive skills such as reading and listening, internal consistency becomes an important criterion in validity claims. Therefore, a number of statistical analysis are carried out using Classical Test Theory (CTT), which allow us to make predictions about the outcomes of testing. Though CTT is regarded as having limited effectiveness due to the fact that the computed indices are group dependent (Alagumalai & Curtis, 2005), it is a practical and valuable means to gather information especially when the test taker population is representative of the intended group of test takers.

The three computations carried out using test scores were: item difficulty index, item discrimination, and reliability.

Each individual item in the reading subtests is scored as either correct or incorrect. For each correct answer the participant students received a 1 and for each incorrect answer they received a 0. The distribution of 1’s and 0’s among the test items provided information on three important aspects of the test/items.

Item facility index (IF) is basically the proportion of test takers who answer an item correctly in a test (Bachman, 2004). The range of the index is between 0.00 and 1.00.

The easier the item the higher is the index value. Hence, an index of 1.00 on an item indicates that all students correctly answered the item and a 0.00 indicates that no one was able to correctly answer the item. The classification of an item as easy (for example, IF>0.80) or difficult (for example, IF<0.40) is arbitrary. As a general rule of thumb, IF values around 0.50 (or a little higher for multiple choice items) are desirable to achieve optimum discrimination between high ability and low ability test takers (Crocker & Algina, 1986). In this study, the acceptable value range was set to 0.40<IF<0.80 through a joint decision by the testing committee and the administration. Item difficulty indices for the items in each reading task were computed using Microsoft Excel (v.2016).

Item discrimination index (ID), which is used to reveal whether responses given to individual items can discriminate between test takers at different proficiency levels (Bachman, 2004; Brown, 2012), was the second analysis carried out on test scores. Although there are a number of ways to calculate ID, point biserial correlation (r_{pbi}) is one of the most commonly used methods (Fulcher & Davidson, 2007). The point biserial correlation evaluates the association between responses to a single item and the test score; that is, it measures to what extent performance on one item in a test is related to the performance on the whole test. This correlation between a single item (a 0 or a 1) and the test score (a continuous variable) can range between -1 and +1, and a value equal to or greater than 0.25 is acceptable (Henning, 1987).

Another measure of ID, in a simpler manner, is to calculate the IF for the top scoring 27% of the participants and bottom scoring 27% of the participants, and subtracting the IF calculated for the top scoring group from the IF of the bottom scoring group (e.g. $IF_{top} - IF_{bottom}=d$) (Brown, 2012). “The higher the value of *d*, the more adequately the item discriminates the higher-scoring from the lower-scoring test takers” (Cohen & Swerdlik, 2009, p. 258).

The third analysis based on the CTT is test reliability, or, internal consistency reliability. There are a number of ways to compute internal consistency. Some of them are based on observed scores in relation to true scores or measurement error, others are based on the correlations of observed scores and true scores or error scores. In this study, reliability is computed as a ratio of item variance and total score variance, which is the most common expression of reliability (Furr & Bacharach, 2008). Coefficient alpha can range between 0 and 1, the latter indicating greater psychometric quality. However, a measure of 1 is not favored because very high alpha levels (for example, above 0.80) may be indication of the scale levels being too narrow or specific. Reliability levels between 0.70 and 0.80 are advised (Furr & Bacharach, 2008). In this study, alpha coefficient calculation was made using IBM SPSS Statistics software (v.24).

This study aimed at validating a reading test by establishing the theoretical basis of how to define a test construct, operationalizing the construct through tasks and items, and investigating which skills/strategies and knowledge components are utilized by test takers when taking the test. Hence, there were three validity concerns: contextual, cognitive and scoring validities. Obviously, there may be other validity concerns such as consequential validity and criterion-related in validation studies. However, the scope of this study was set at investigating and presenting validity evidence on three aspects of test development mentioned above.

CHAPTER 4

RESULTS

4.1 Introduction

This study aimed to investigate the stages of test development as specified in the sociocognitive framework and to provide evidence related to the context, cognitive and scoring validities of a reading test. To investigate the context validity of the test, first, a local needs analysis study was reviewed and its implications for the test were established. Then, a reading model from the literature was used as a framework to define reading ability as a test construct. Finally, the reading construct was operationalized and the results were presented in the test specifications document (APPENDIX E).

In the next step, cognitive validity was investigated through introspection and retrospection of the test taking processes by two different participant groups. The skills and strategies the participants used while responding to test items provided information on whether the cognitive processes reflect the activities specified in the reading model used in the development of the test. Finally, scoring validity was sought through statistical analysis of test scores. The present chapter provides the results of data analysis on these three aspects of test validation.

4.2 RQ1: How is Academic Reading Ability Conceptualized and Operationalized as a Test Construct?

4.2.1 Defining reading ability: Historical perspective. The literature is abounded with different approaches to and definitions of reading ability, both in L1 and L2. Urquhart and Weir provide one: “Reading is the process of receiving and interpreting information encoded in language form via the medium of print” (1998, p. 22). Through this straightforward-looking definition one can draw a number of conclusions: first, it is understood that there is some kind of interaction between the reader and the printed text. This interaction helps the reader to construct meaning using information from the test and their own knowledge and skills (Grabe, 1991). Secondly, a number of processes are activated during *receiving and interpreting information.* Lower level processes such as word recognition, syntactic parsing, or higher level processes such as inferencing come into play (Grabe, 2009). Finally, the reader makes use of their knowledge of the language (e.g. writing system, or vocabulary), knowledge of the topic of the text, or knowledge of the world to make meaning of the text.

Apparently, many skills, knowledge basis and components of cognitive ability are involved in reading. Not surprisingly, reading has been studied by cognitive psychologists, whose work on this subject produced various models of reading, some of which have been used widely in language teaching (Urquhart & Weir, 1998). Those models mainly fell into two categories: the process and componential views of reading. Process models of reading focused on describing the actual cognitive processes that take place during reading (Urquhart & Weir, 1998) whereas componential models attempted to describe the subskills that are believed to underlie reading ability (van Steensel, Oostdam, & van Gelderen, 2013).

4.2.1.1 Process models. There are three main approaches in explicating the process of reading. The most commonly encountered are the bottom-up and top-down models of reading, which were originally used in computer science to distinguish between *data-driven* and *knowledge-driven* processes (Field, 1999). A third model was introduced more recently: the interactive information processing model.

4.2.1.1.1 Bottom-up models. Reading as a bottom-up process model was advocated, among others, by Gough (1972) who suggested that reading starts with recognition of letters, then phonemes, and words; in other words, the reader decodes printed text starting from the smallest unit moving progressively to larger units. During decoding syntactic and semantic rules are activated to understand sentence meaning.

In this model, the focus is on the processing of the constituents of texts in a sequential order (Moore, Morton, & Price, 2007; Treiman, 2017; Urquhart & Weir, 1998). It has been claimed that readers with low language proficiency rely more on decoding skills, and parsing sentences into constituent parts, which suggest the dominance of a bottom-up approach in reading comprehension for them (Treiman, 2017; Verhoeven, Reitsma, & Siegel, 2011).

A number of pitfalls in explaining the process of reading using only this approach were expressed. One of them was that if there were actually a strict sequence of recognition starting from the letter then it would take a longer time to recognize a word, which is not the case (Urquhart & Weir, 1998).

Another argument against the bottom-up model was that readers use syntactic information to decipher word meaning when there is ambiguity, and this strategy points to a different direction of processing rather than the bottom-up model (Urquhart & Weir, 1998). Urquhart and Weir argue that in bottom-up models there is logical inconsistency: if the reader needs to understand all the words before she can understand the meaning residing in the sentence then how does she know when and where to stop processing words?

4.2.1.1.2 Top-down models. As opposed to bottom-up models that start from the smallest text unit moving gradually to the whole of text, top-down models start from the whole text and move down to smaller units. The context and domain specific knowledge contribute to the understanding and constructing of the meaning in the text.

One approach in top-down model of reading is to explain this process as *hypothesis verification, "*whereby the readers use selected data from the text to confirm their guesses" (Urquhart & Weir, 1998, p. 42). The hypothesis may start from context previously provided in the text, from the reader's own knowledge of the topic, and knowledge of the types of texts that are presented in a particular genre of book (Pearson & Kamil, 1978). Goodman (1997) is known to advocate top-down processing approach within his 'whole language' theory by arguing that "readers bring a great deal of knowledge, expectations, assumptions and questions to the text and, given a basic understanding of the vocabulary, they continue reading as long as the text confirms their expectations" (Gamboa-González, 2017). In this model, one of the prominent difficulties is the challenge of dealing with the whole text in starting to read.

4.2.1.1.3 Interactive models. Continuing research on reading revealed that neither bottom-up nor top-down models could satisfactorily explain the reading process and comprehension on their own, and that there is an interactive process between these two types of models. There were many proponents of this view of reading (Carrell, 1988; Hudson, 1998; Stanovich, 1980; Verhoeven et al., 2011). Among those, McClelland and Rumelhart's (1981) model is well-known, which was based on an earlier work by Rumelhart (1977). Known as **Interactive Activation Model,** it was based on three assumptions:

> (i) There are a series of levels of processing of letters and words as well as higher levels of processing (top-down) providing contextual input to the word level.
> (ii) There is parallel processing during reading. That is, while the reader processes at the letter level she also processes at the word level and at the text level.
> (iii) Perception is an interactive process. In other words, top-down and bottom processes work simultaneously to help the reader perceive the text.

Today, it has been widely recognized that reading process is a combination of both the processing of visual information (bottom-up) and world knowledge put in use through the text (top-down) (Khalifa and Weir, 2009), and that these processes may take place simultaneously or in an integrated manner (Faerch & Kasper, 1986; Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003). The main argument in this approach is the utilization of information coming from different sources simultaneously (i.e., bi-directionality in reading processes) rather than sequentially.

4.2.1.2 Componential models. Componential models of reading assume that reading ability can be analyzed and therefore tested through its components. Dividing reading skill into sub processes relied mainly on factorial analysis of test takers' performances (Johnston, 1981). In this method, different reading tasks or items were expected to load on a limited number of factors that explain the underlying variables of reading.

Studies on this divisibility approach go back to 1944 when Davis claimed that he found eight separate subskills of reading. In a later study, he amended this number to five ((i) memory for word meanings, (ii) inferencing, (iii) following passage structure, (iv) recognizing a writer's purpose, attitude, tone and mood, and (v) finding answers to questions asked explicitly or in paraphrase) and argued that comprehension of a reading text is not a unitary skill or process (1968, in Khalifa and Weir, 2009). Following Davis, Spearitt (1972) conducted analysis on the same data set but found only the first four factors. A third analysis of the same data set by Thorndike (1973) ended with evidence for two factors only: general reading comprehension and word knowledge.

The assumption that reading could be analyzed through its constituent competencies was not favored by all. There was evidence for both divisible and non-divisible views of reading (Khalifa & Weir, 2009). Rosenshine (1980), for example, examined data from previous studies and eventually claimed not to have found any evidence of the existence of discrete skills in reading. Another counter argument to the divisibility view came from Alderson (2000). He claimed that the whole reading process is integrated and that the skills needed to understand a text cannot be identified empirically.

In the presence of studies with conflicting results, Khalifa and Weir (2009) argued that sampling, method of analysis and the tasks used in the test affected the outcomes of analyses. They also maintained that the assumptions of the statistical approach were flawed because the analysis focused on the factors that could be statistically shown to contribute to performance on the tests rather than the actual processes that a reader carried out in real life. Hence, they argued that the data examined were a measure of success on a test rather than successful reading.

At the same time with the use of the factorial approach to identify components of reading, **a skills approach in teaching** was widely favored. This approach divided reading into skills and subskills with a focus on behavioral outcomes in instruction. Khalifa and Weir (2009) argued that designing taxonomies with subskills for reading might have initiated the orientation towards a communicative syllabi and the need to teach language components in smaller chunks. It was generally agreed that devising taxonomies and identifying microskills was based on informed intuition rather than empirical research (Anderson & Lukmani, 1989; Khalifa & Weir, 2009).

Munby (1978), among others, was one of the most influential figures in advocating the subskills approach. He identified 266 microskills which he believed help the readers to understand the texts. Some of the more important microskills identified by Munby are:

-Understanding explicitly stated information

-Understanding information when not explicitly stated

-Understanding conceptual meaning

-Understanding the communicative value of sentences

-Understanding the relations within the sentence

-Understanding relations between parts of text through lexical cohesion devices

Though dividing skills into subskills and strategies was widely accepted in the teaching of second language practice due to the practicality it offered in material development and syllabus design, Khalifa and Weir (2009) draw attention to the inconsistencies in the terminology used in describing reading processes. This shortcoming may be due to the fact that the subskills approach is more organizationally driven rather than theoretical, as mentioned earlier.

4.2.1.3 A cognitive processing model in reading. The present study employed a model of reading that is based on the cognitive processing approach. As Khalifa and Weir (2009) argued, both the factorial and subskills approaches overlooked the major role of the test taker in a reading process. They posited that there was very little reference to the actual cognitive processes that take place in the minds of the readers when they undertake a reading task. They argued, therefore, that both the context and the cognitive processes should be considered in defining the construct of reading in a test. Following this idea and in the endeavor to define a more accurate model of reading as a cognitive process in the local context, I made use of a cognitive processing model to define reading ability as a construct for a reading test. This model is taken from works by H. Khalifa and C. Weir (Khalifa & Weir, 2009; Weir & Khalifa, 2008a) who utilized Just and Carpenter's (1980) and Kintsch and van Dijk's (1978) earlier works to develop their own (Figure 9).

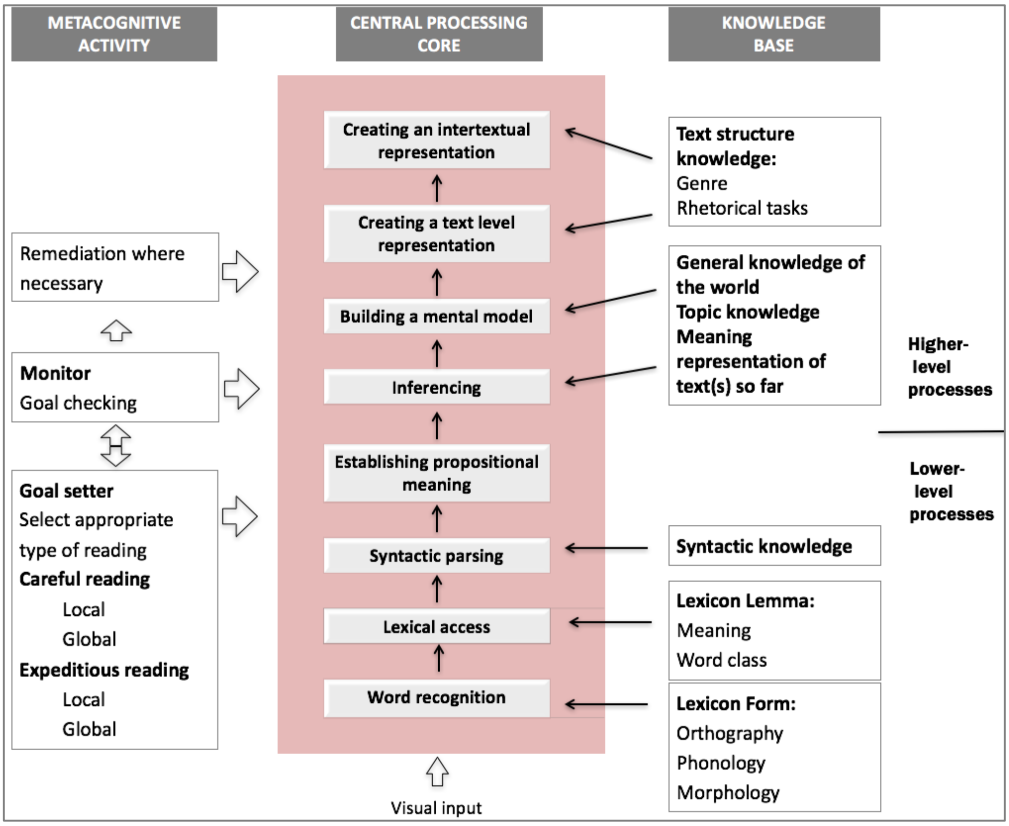


Figure 9 Khalifa and Weir's (2009) model of cognitive processing in reading

Khalifa and Weir's model of cognitive processing in reading consists of three interlinked parts:

- **Metacognitive Activity** that defines the type of activities that the reader carries out.
- **Central Processing Core** that includes elements initiated by the activities carried out in the Metacognitive Activity.
- **Knowledge Base** which refers to the types of knowledge that the reader brings into the reading process.

Metacognitive Activity includes three separate parts: **Goal Setter, Monitor** and **Remediation**. Goal setter is where decisions about the purpose of reading are given. Depending on the type of reading, i.e. expeditious or careful reading, some processes in the central processing core are activated. As reading activity is an interaction of top-down and bottom-up processes, the processes in the **Central Processing Core** and the **Knowledge Base** are activated depending on the purpose of reading.

Goal Setter is the agent that determines the purpose for reading and decides what type of reading will be carried out to achieve that purpose. Two types of reading purposes are specified: careful and expeditious reading which are carried out either at the local or global level. These purposes and reading dimension were previously theorized in Urquhart and Weir (1998) as a four-cell matrix (Table 13). This matrix provides the link between reading types and the processes carried out in the Central Processing Core.

Table 13 Urquhart and Weir's (1998) matrix of reading types

	Local	**Global**
Careful	Understanding syntactic structural sentence and clause. Understanding lexical and/or grammatical cohesion. Understanding lexis/deducing meaning of lexical items from morphology and context.	Reading carefully to establish accurate comprehension of the explicitly stated main ideas the author wishes to convey; propositional inferencing.
Expeditious	Scanning to locate specific information; symbol or group of symbols; names, dates, figures or words.	Skimming quickly to establish discourse topic and main ideas. Search reading to locate quickly and understand information relevant to predetermined needs.

Careful local reading is primarily bottom-up processing in order to decode the propositional meaning. There is no need for building meaning representation at a higher level. The level of processing is limited to a clause or a sentence, and the meaning is derived from this local context. The following sentence is from one of the reading tasks: *There are some well-known methods to find out other people's passwords.* In this example, meaning is derived from the relation between the sentence and what it describes in a real or imaginary world. As such, it can be gauged to be true or false. This type of reading is carried out when the meaning resides within a clause or a sentence.

Careful global reading, as the name implies, is related to understanding the text as a whole by paying attention to parts of the text that the writer considers important, and to establish a macro structure of the text through the information received (Khalifa and Weir, 2009). In this type of reading, most of the components of the Central Processing Core are initiated. The reader processes the whole of the text (which may be a couple of sentence, a paragraph, or a whole text) through bottom-up and top-down processes: she starts decoding, but also uses her knowledge base after establishing propositional meaning of parts of the text, to make inferences or to connect with prior knowledge on the topic.

This is very similar to real world reading where the reader combines information from a number of sources; for example, in freshman courses at METU, project assignments require students to read different sources on the subject and synthesize information in their assignments.

The following passage is from one of the reading tasks:

> *Stanford University, which is listed among the top ten universities in all over the world, suggests to its users a very simple password building procedure, the result of which is also very simple. They suggest selecting four simple words and concatenating them into a passphrase. Their example is orange+eagle+key+shoe. Calculating with quite a big dictionary of eight thousand basic words, we get that it would need 68 minutes to crack such a password. If the dictionary contained only two thousand words, the time requirement would be only 16 seconds.*

In this passage, in order to understand the writer's opinion of Stanford University's suggestion regarding password building, the reader needs to read the whole passage carefully. Only by forming a relation between sentences and reading between lines can one understand that the writer of the text did not find the suggestion reliable. This skill, inferencing, is considered a higher level reading skill as the reader tries to extract implicit meaning from the text.

Expeditious reading is defined as quick, selective and efficient reading in order to arrive at the desired information (Urquhart and Weir, 1998). Three reading strategies can be grouped under expeditious reading: Skimming, scanning and search reading.

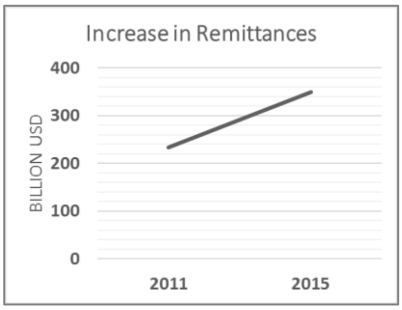
Skimming is generally defined as reading quickly to get the gist or the main idea of a text. There are certain strategies applied in getting the gist of a text:

- reading quickly and selectively
- reading the first and last paragraphs of a text
- reading the first and last sentences of a paragraph
- reading the titles and subtitles
- paying attention to non-textual information (charts, pictures, etc.)

The reader may employ one or more of the strategies given above to be able to understand the macro structure of a text. However, this type of reading does not allow the reader to create a detailed representation of a text in their minds (Khalifa and Weir, 2009). For example, details, or implicit meanings cannot be understood through skimming. Although skimming involves some local reading at sentence level, in general it is considered **expeditious global reading**. In the following sample paragraph (page 80), in order to establish the macro structure, the reader can approach it by reading the title of the line graph and by skimming through the text and reading words randomly. The gist of the text can be understood to be about remittances sent by immigrants, and that the remittances were on the increase from 2011 to 2015. ***Scanning*** is also selective reading but to achieve a very specific reading goal. The reader looks for some very specific information such as dates, figures, names, etc. The activity in the Central

Processing Core is mainly decoding: word recognition. There is no need to make meaning at sentence or text level; therefore, it is mainly **expeditious local reading**.

> *Mass migration has produced a giant worldwide economy all its own, which has accelerated so fast during the past few years that the figures have astounded the experts. This year, remittances — the cash that migrants send home — through banks is set to exceed $232 billion, nearly 60% higher than the number just four years ago. Of that, about $166.9 billion goes to poor countries. In many of those countries, the money from migrants has now overshot exports, and exceeds direct foreign aid from other governments since there are many people sending 40% of their income in remittances. Indeed, many experts believe that the true figure for remittances this year is probably closer to $350 billion, since migrants are estimated to send one-third of their money using unofficial methods, including taking it home by hand. That money is never reported to tax officials, and appears on no records.*



In the reading passage given on page 79, a question such as *How long would it take to crack a password using a big dictionary*? would be answered through scanning for a figure (*68 minutes*) and crosschecked with keywords (*big dictionary*).

Search reading is slightly different from skimming and scanning as it involves reading at both local and global level. According to Urquhart and Weir (1998) the reader searches for a pre-determined topic by sampling the text at different levels (words, topic sentences, introductory paragraphs, etc.) in order to extract information. Different from skimming, the reader does not establish a macro structure of the text, but is involved in getting information on a specific topic, the location of which is determined by quick reading. With relation to the Central Processing Core, the reader looks for words in the text belonging to the same semantic field of the topic she is searching for. Once it is located, the reader reads carefully to establish propositional meaning, or to make inferences. There is no need to create textual or intertextual representation, but world knowledge and topic knowledge may be used to make meaning of the text. If the information sought can be found within a sentence it is **search reading local**, if it involves comprehension of more than one sentence, it is **search reading global**.

Part of a text from the search reading task in Reading Test V2 is given on page 83. In order to answer a specific question such as *Why is there a specific layout for the chambers?* the reader needs to identify the location of this information (II. Description of the facility) by skimming the text and reading subtitles. Afterwards, the reader reads carefully paying attention to the keywords (*Why* – in the question, *purposeful* in the related sentence) to answer the question (It serves as a security measure against an explosion.).

SVALBARD SEED VAULT

I. Introduction

The general public is well aware of the threat of extinction to animal species, far fewer are aware of the risk of crop extinction. With whales or tigers or polar bears," you can look at them in the eye and you can be very empathetic. But you can't do that with a wheat variety or carrot variety”. The history of Svalbard seed vault starts as early as 1983. Like other big projects, it’s been a long and not very easy journey. Preserving seed from food plants is an absolutely essential part of the work of preserving the world’s biodiversity, adapting to climate change and global warming and eventually ensuring food for the world’s population for the foreseeable future. The foundation of a global central seed bank for the world’s seeds (primarily of food plants) has therefore long been an issue and Svalbard Global Seed Vault was a step in this direction. In 1989, the International Board for Plant Genetic Resources (IBPGR) started surveying the relevant alternative sites in Svalbard. Norway offered to take care of the actual construction of the vault, while the Food and Agriculture Organization (FAO) and IBPGR would take care of the administrative operating costs through the creation of a fund based on capital from external donors.

II. Description of the facility

This Seed Vault lies about 1 kilometer from Longyearbyen Airport, at about 130 meters above sea level and consists entirely of an underground facility, blasted out of the permafrost (at about minus 3-4 degrees Celsius). The facility is designed to have an almost “endless” lifetime. The location takes into account all known scenarios for rising sea level caused by global climate changes. The facility has also been located so deep inside the mountain that any possible changes to Svalbard’s climate, which we know about today, will not affect the efficacy of the permafrost. This will be a temporary temperature back up in the event of technical failure, such as loss of power supplies for a period.

The facility consists of three separate underground chambers. The layout of these chambers is purposeful. None of them are in a direct line. Instead, the workers have carved out a concave indentation in the rock. This serves as a security measure against an explosion. The chambers, each of which have the capacity to store 1,5 million different seed samples, will have storage shelving for pre packed examples of food seeds from the depositors.

The monitor in Metacognitive Activity checks whether the reading activity carried out is consistent with the goals set. This activity is called Goal Checking. It takes place during decoding and meaning building. It has been argued that skilled readers check their comprehension of the text regularly by forming meaningful links between sentences as they read, and fill in the missing parts by making inferences. Goal checking requires the reader to identify what they do and do not understand while reading and where the difficulty lies. If during that process, the reader notices that they fail to understand the text, remediation comes into play. Several strategies can be used during remediation: the reader might adjust their reading speed to fit the difficulty level of the text, or try to translate or paraphrase the parts that are found to be difficult to understand. Looking back or forward in the text might also be used to resolve the difficulty.

The Central Processing Core comprises eight processes sequenced hierarchically:

- Word recognition
- Lexical access
- Syntactic parsing
- Establishing propositional meaning
- Inferencing
- Building a mental model
- Creating a text level representation
- Creating an intertextual representation

Word recognition is matching a word form in a text with the representation of an orthographic form known to the reader (Khalifa and Weir, 2009). In the case of experienced readers, the matching of the form of a word with the mental representation is automatic. However, if the reader is not fluent in the language she may need to use much of her cognitive skills in decoding words, and not be able to build meaning from the text. As such, item writers are advised to make sure to control the vocabulary range of the texts given to inexperienced readers so that their resources will not be exhausted at this level.

Lexical access is about the orthographic, phonological and sometimes morphological mental representations of a word. At this level, the reader matches the form of a word with its mental representation as well as the meaning. The more frequent words are presented to the reader, the more quickly they are matched with words in the readers' mental vocabulary (Khalifa and Weir, 2009).

Syntactic parsing is about the grammatical structure of the text. After deciphering the word form and meaning, the reader makes sense of larger units such as a clause or a sentence through her knowledge of the syntactic structure of the language. It has been posited that in assessment it is important to present examinees with syntactic categories in accordance with their level of knowledge of syntax, morphology and other grammatical elements (Khalifa and Weir, 2009).

Establishing propositional meaning refers to the literal interpretation of the smallest meaningful unit, a clause or a sentence, without any inferencing. Establishing the literal meaning of a clause or sentence does not require any higher order interpretive factor. It is simply decoding of the printed text and deriving the propositional meaning.

The processes to this point are called lower level processes. The following stages of processing belong to higher order processes that require the reader to create meaning above the sentential level by making use of knowledge bases such as topic knowledge, world knowledge and text structure knowledge.

Inferencing is a higher order process. Khalifa and Weir (2009) posit that inferencing is a creative process because the reader needs to fill in the gaps between ideas by adding information that is not explicitly stated. They also state that inferencing does not always take place at the sentence level but also at word level in which case the reader needs to guess the meaning of the word by using contextual clues.

Building a mental model, the stage after inferencing, refers to the consistent adding up of new information onto what has been read before. Field (2004) maintains that

> [i]ncoming information has to be related to what has gone before, so as to ensure that it contributes to the developing representation of the text in a way that is consistent, meaningful and relevant (p. 241).

As the reader is engaged with the text, she may update, or modify the mental model she builds with the new information she receives from the text through monitoring.

Creating a text level representation is a stage where the reader understands the hierarchical structure of the text, and identifies parts of the text that are significantly related to the main idea(s) of the text. At this level of processing, the reader understands the discourse structure of the text and distinguishes main ideas of the text from others (Khalifa and Weir, 2009).

Creating an intertextual representation refers to comprehension of multiple texts by creating a macro-structural organization in order to connect representations of those texts meaningfully (Lacroix, 1999, in Khalifa and Weir, 2009).

The next section of the model is called **Knowledge Base.** This section of the model reveals the types of knowledge activated with relation to the cognitive processes that are employed. Some knowledge types have been grouped together as they are usually activated simultaneously. The types of knowledge presented in the model are:

- Lexicon (Form and Meaning)
- Syntactic knowledge
- General knowledge of the world / Topic knowledge / Meaning representation of text so far
- Text structure knowledge (Genre and Rhetorical tasks)

Lexicon refers the list of words of a language or communication system (Zeevat, Grimm, Hogeweg, Lestrade, & Smith, 2017). Knowledge of lexicon in this reading model contains information regarding the form of words, i.e. the orthography, phonology and morphology of words, and information regarding the meaning of the word and the word class, i.e. whether it is a noun, verb, adjective or adverb.

As the reader gets the visual input from a text, the processes that are activated initially are word recognition and lexical access. These two processes rely on the information received from the mental lexicon of the reader. During word recognition, the form of a written word is matched with a representation of the orthographic form. Experienced readers may make this connection automatically, whereas for less experienced readers this process could be complicated (Oakhill & Garnham, 1988). During lexical access, information about a word's form and meaning is retrieved (Field, 2004). Words that the reader encounter frequently would be more quickly identified, which suggests that in test construction the amount of frequent and less frequent words to be included in the texts need to be considered with regard to the level of proficiency of the test takers. It is expected that knowledge of less frequent words will increase as a reader becomes more proficient in language use.

Syntactic knowledge is the knowledge of how words can be combined in meaningful sentences, phrases, or utterances. After accessing the lexicon to receive information about the form and meaning of words, the reader puts words together to make phrases, and then creates the larger units of clause and sentence to understand the message. Establishing the propositional meaning of the sentence takes place at this level.

At the level of **General knowledge of the world / Topic knowledge / Meaning representation of text so far,** more knowledge is added to the propositional meaning of the sentence to make it meaningful in the context it appears. Meaning representation of text involves the macro-structure of text which is formed as cohesive links between text are formed. The reader uses world knowledge or knowledge of topic to "judge the coherence and consistency of what has been understood when it is integrated into the ongoing meaning representation" (Khalifa & Weir, 2009, p. 52).

At **Text structure knowledge** level, the reader has already established the discourse-level structure of the text. She then determines, through the use of knowledge of genre and rhetorical tasks, how the text is structured and which parts of the text are important for the purpose of the writer. Through the knowledge of discourse, the reader also identifies the macro level relationships between ideas.

4.2.2 Freshman students' communicative needs. After establishing the conceptual basis of reading ability using a cognitive processing reading model that is applicable in academic contexts, I now present a brief review of a needs analysis project carried out at METU. This review brings to light specific reading requirements in the first year of undergraduate programs at the five faculties (Engineering, Education, Architecture, Arts and Science, Economic and Administrative Sciences) at METU. The next step is amending the cognitive processing reading model in accordance with the contextual needs, which finalizes the localization of the reading model and establishes the test construct of reading in an EFL context.

4.2.2.1 Background. The needs analysis study was part of a larger project that sought to investigate the curricular activities of the two departments, Department of Basic English (DBE) and Modern Languages Department (MLD), under the School of Foreign Languages (SFL). DBE teaches English to newly registered undergraduate students whose level of English proficiency is not sufficient to study at an academic program at METU. MLD also provides English language instruction, besides other languages, but their courses are for freshman, junior and sophomore students.

The aim of the above mentioned project was to maintain high standards in SFL's activities and improve school effectiveness by

1) reviewing the current situation at the macro and micro level, and assessing how effectively the school can respond to local and global transformations,
2) analyzing the target language needs of the students and renewing the curricular programs in line with the findings,
3) defining English language proficiency in the light of our students' future language needs
4) developing a language proficiency test that conforms with systematic test development procedures and validity theories.

4.2.2.2. The design and implementation of the project. The project was carried out by the two coordinators of the Research and Development Unit under the SFL, one instructor from DBE, and one from the MLD. Each coordinator was responsible for investigating students' needs in their respective domain. As the coordinator on the DBE side of the project, I started working on the project in 2013, and designed a research study that investigated 1) the effectiveness of the programs offered at the DBE at five different levels (beginner, elementary, intermediate, upper-intermediate, and the repeat group), and 2) DBE students' language and learning needs in the first year of their subject studies. The stakeholders included people from the DBE, the MLD and the faculties. The instructors at the MLD and the faculties had first-hand experience of the difficulties undergraduate students face during their freshman year. In the same vein, students from the DBE and first year students from the faculties were invited to participate in the study. Data collection was carried out through questionnaires, focus group and one-to-one interviews with the help of two other colleagues. In the summer of 2014, qualitative and quantitative analyses were completed. The summary of the research design is given in Table 14.

Table 14 Research design of the SFL project

Focus	Purpose	Participants	Procedure	Evaluation & Analysis
Curriculum	To find out whether the planned & implemented curricula are compatible	R & D	Document Analysis	Discourse
	To evaluate the efficiency and effectiveness of the goals and objectives of the implemented curriculum	DBE Instructors	Interview Questionnaire	QUAL QUAN & QUAL
		DBE Students	Questionnaire Focus Group	QUAN QUAL
		DBE Instructors	Interview Questionnaire	QUAN & QUAL
		DBE Students	Questionnaire Focus Group	QUAN & QUAL
	To find out whether the implemented curriculum efficiently prepares DBE students for their departmental studies	Freshmen	Questionnaire	QUAN & QUAL
		DML Faculty Freshman	Focus Group Interview Questionnaire	QUAL QUAL QUAN
		DML Freshman Faculty	Focus Group Questionnaire Interview	QUAL QUAN & QUAL QUAL
		Freshman Students	Questionnaire	QUAN
Materials	To find out whether commercial and in-house materials effectively address DBE students' needs & interests	DBE Students DBE Ins.	Questionnaire Focus Group Interviews	QUAL & QUAN QUAL

4.2.2.3 Defining communicative needs. The aim of the needs analysis study (SFL, 2015) was to investigate the communicative tasks students needed to carry out in the first year of their subject studies. Ultimately, the DBE curriculum and the syllabi, and the content of the language proficiency exam were planned to be based on the findings of the needs analysis study (Figure 10).

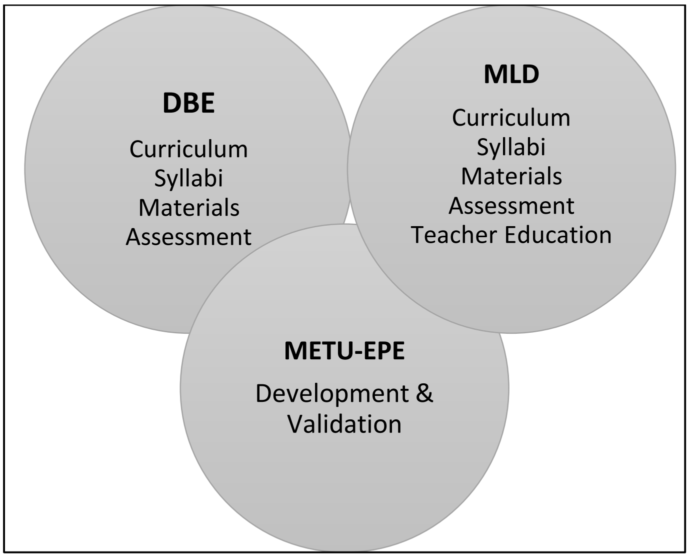


Figure 10 Domains of the SFL project

4.2.2.3.1 Data sources: Interviews. The interviews conducted with the faculty members, freshman students, and DBE instructors were the major sources of data. The faculty interview questions focused on the actual course requirements such as the type and density of reading to be carried out as preparation for the courses each week, the types of assessment batteries, test tasks, and assignments given to students. Moreover, faculty professors' and MLD instructors' observations on the strengths and weaknesses of freshman students in carrying out course requirements were investigated through the interviews.

DBE Instructors were interviewed to reveal the effectiveness of the curriculum in addressing students' needs, and how to improve instruction. Their views on the importance of each language reading skill (reading, listening, writing, speaking) were also investigated.

Focus group interviews were carried out with DBE students to reveal areas where they find instruction at the DBE effective/ineffective, and their motivation levels regarding the learning of each skill.

4.2.2.3.2 Data sources: Documents. A second set of data came from the documents collected from the web such as the syllabi, assignments and course books announced on the department web sites. Content analysis of the documents informed me about the weight and variety of the workload of freshman students (reading assignments, project assignments, lab assignments, etc.), and the types of texts (genre, style, complexity, etc.) that the students were expected to deal with.

4.2.2.3.3 Data sources: Questionnaires. Questionnaires were used to collect large amounts of data from DBE and freshman students on various aspects of teaching and learning. For example, freshman students were asked to rate the importance of the communicative tasks (e.g. group work, discussions, background reading as preparation for lectures, etc.) for achievement in their respective programs. They were also asked to rate the effectiveness of the instruction they had received from the DBE. The participants were provided space to comment further on the questions asked.

DBE students were administered a detailed questionnaire investigating their views on the effectiveness of the instruction they were receiving at the time, improvements they could suggest, the ranking of the language skills in importance for them, etc. Their views on the assessment system and the materials were also investigated.

A questionnaire very similar to that of the DBE students was administered to the DBE instructors – for the purpose of comparison of opinions.

The data were analyzed quantitatively, and where possible, comparisons were made between the responses.

4.2.2.4 Data analysis. As the language proficiency exam is given to all students independent of their field of study, one of the main aims while analyzing the data was to identify the language needs that were common to all disciplines.

For a better understanding of the common approaches and methods used in teaching in various disciplines, the data was categorized into four major disciplines following Biglan's (1973) grouping of the different scientific areas in accordance with their epistemological origins and research methodologies.

These four categories were **Hard Applied Sciences (HAS), Hard Pure Sciences (HPS), Soft Applied Sciences (SAS)** and **Soft Pure Sciences (SPS)** (Table 15). Nonetheless, the boundaries between the disciplines are not solid and sometimes there may be overlaps between them.

Table 15 Faculties and student distribution (2014)

	PURE		***APPLIED***	
	Field	***Student %***		***Student %***
HARD	Biology Chemistry Mathematics Physics Statistics	14	Aerospace Engineering Chemical Engineering Civil Engineering Computer Engineering Electrical and Electronic Engineering Environmental Engineering Food Engineering Geological Engineering Industrial Engineering Mechanical Engineering Metallurgical and Materials Engineering Mining Engineering Petroleum and Natural Gas Engineering	47
SOFT	Economics History Philosophy Political Science & Public Adm. Psychology Sociology	22	Architecture Business Administration City and Regional Planning Computer Ed. and Ins. Technology Elementary Education Foreign Language Education Industrial Design International Relations Secondary Education	17

4.2.2.5 Findings.

4.2.2.5.1 Importance of reading. One major finding of the needs analysis study was the apparent emphasis on the importance of the reading skill among others in all disciplines. The instructors offering the freshman courses established that reading, and especially critical reading of academic texts, was a prerequisite for success in academic programs. Shih (1992) states that reading for academic studies involves critically reacting to the text, recalling what is read (both the main points and details), synthesizing related information from readings and lectures. Reading, in this sense, entails such skills as concentrating, planning, critically analyzing, synthesizing and evaluating. Furthermore, the instructors of the DBE maintained that the teaching of the reading skill should be prioritized as they considered it to be an essential academic skill. According to both the instructors in the faculties and in the pre-sessional language school, reading is the most important skill in learning an academic subject. Although the data gathered from the students at the DBE and the freshman courses revealed their preference for speaking (DBE) and listening (freshman) as their first choice, reading was perceived as a very important academic skill by all freshman students (HAS students: 90%, HPS students: 86%, SAS students: 91%, SPS students: 97%).

4.2.2.5.2 Types of reading texts. After establishing reading as a major and critical skill in academic achievement, the next step was to identify the types of texts students read. In almost all disciplines, students were required to follow course books. As a preparation for the lectures, they were expected to do background reading each week. The weekly reading load ranged from a few pages to about 40 pages including chapters from course books, newspaper articles, scientific articles, and technical texts from reference books or the Internet. The majority of the course books were of foreign origin written for college-level students. The types of texts in such resources were expository, descriptive, and sometimes argumentative. Expository type of texts dominated in all fields; however, argumentative texts such as comments and books reviews were also among required readings.

In terms of the length of texts, the analysis revealed that the texts assigned to freshman students range between 20-40 pages per chapter/week. This finding pointed to the need to use longer texts in assessing reading ability so that the test tasks better represent the actual tasks.

4.2.2.5.3 Vocabulary. All participants agreed upon the fact that knowledge of vocabulary is very important for comprehension, and that most DBE students lacked this knowledge. The instructors expressed their wish that the students were equipped with general academic vocabulary before they started their degree programs.

Although the concept of academic vocabulary is vague, there are helpful resources in the literature that categorize lexis according to frequency of use in general and academic settings. Two of those resources are the New General Service List (NGSL) (Browne et al., 2013b) and the New Academic Word List (NAWL) (Browne et al., 2013a), which consist of words relevant to academic study. The NGSL consists of 2800 words, and the NAWL consists of 963 headwords (that is, time and plural inflections of words included), which cover 92% of academic texts, as stated by the authors. The lists were derived from a 288 million-word corpus of academic texts from the United Kingdom and United States covering a wide range of academic disciplines.

In order to ensure that the NGSL and NAWL sufficiently address the vocabulary needs of students at METU, randomly chosen chapters from one or two course books used in various programs (branches of engineering, sociology, history, philosophy, education, international relations, economy) were submitted to a vocabulary analyzer software (Compleat Lexical Tutor: http://www.lextutor.ca/vp/). The results were very similar to those announced by the authors of the NGLS and NAWL. Hence, it was decided that these two lists could serve as the basic level of lexical range students need to master.

4.2.2.5.4 Describing the reading skill. According to the findings of the needs analysis study, the purposes of reading at freshman level were categorized using Council of Europe's language framework (2009) as:

1) Reading for information and argument: The goal is to comprehend the main ideas and other essential information, stated either explicitly or implicitly. One should also be able to identify the arguments in a text. This type of reading is carried out when the students read to learn as preparation for an exam or project. The types of texts relevant for the students are course books, books, articles, instructors' notes, slides, own notes etc.

2) Reading for orientation: The goal is to quickly locate information on a predetermined topic in long and complex texts, and read only the parts that are needed. This type of reading is also helpful in establishing whether the content of a text is relevant and whether a specific part of the text needs careful study. The types of texts used are books, articles, other relevant texts from the internet, technical reports, etc.

3) Reading instructions: The goal is to understand everything in detail. The texts are short, from a sentence to a few sentences. These may be exam instructions, exam questions, other instructions relevant to class work (report preparation instructions, project assignments, etc.)

4.2.2.6 The new conceptual model. A viable reading construct for this setting is based on the four-cell matrix of reading types proposed by Urquhart and Weir (1998) and elaborated by Khalifa and Weir (2009). Here, the reading types, i.e., careful/expeditious and the reading dimensions local/global are the two continua along which the three categories of reading activities specified for undergraduate students are positioned.

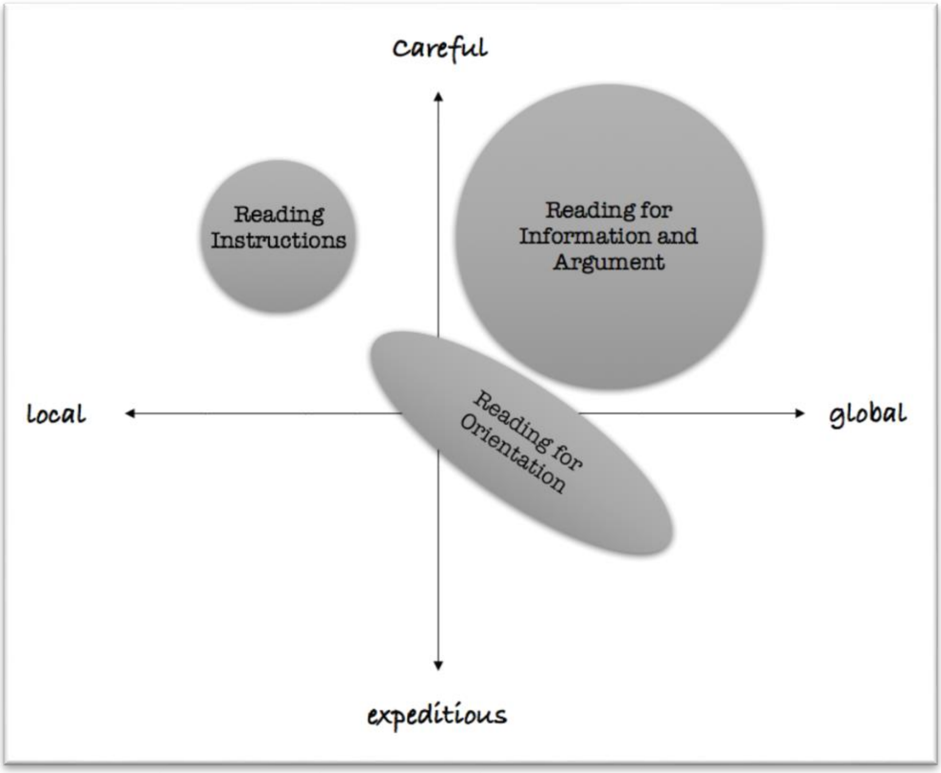


Figure 11 Conceptualized reading model according to reading purposes and types

In this model reading construct is defined in two dimensions and with reader purposes: **reading instructions, reading for information and argument,** and **reading for orientation**.

Reading instructions involves reading questions or scenarios in the exam papers, reading instructions for assignment preparation, reading announcements, reading essential information in the syllabi regarding course requirements, assessment procedures, etc. In this case, reading is carried out mainly at careful local level, but occasionally may require careful global reading as well. At the local level, the student decodes the text to establish the basic meaning of the sentence. Some inferencing may also be necessary to build a mental model. As the meaning resides usually at the sentence level, there is no need to integrate pieces of information to build a larger textual meaning representation. Regarding the reading model (Figure 9), while reading instructions decoding (word recognition, lexical access and syntactic parsing) and establishing propositional meaning at the sentence level take place.

Reading for orientation mainly requires global expeditious and global careful reading. This type of reading is necessary to search for some specific information on a predetermined topic or to understand the gist of a text. Students read for orientation in preparation for a course: they may skim through reading assignments before class meetings in order to understand the main ideas in the course material. The attention is on parts of a text that have macro-propositional character. In order to understand whether a proposition has such a value the reader uses general/world knowledge to make guesses. However, knowledge of genre and how texts in different genres are structured, also help the reader to choose possible positions of the macro-propositions. For instance, in a formal argumentative text, the reader would find the argument of the text within the thesis statement generally placed in the introductory first paragraph of an essay.

Reading for orientation may also involve scanning. In scanning, the reader's aim is to access very specific information. She searches for keywords, numbers, dates, etc. Khalifa and Weir (2009) define it as "a perceptual recognition process which is form based and (which) relies on accurate decoding of a word or string of words" (p.59). The part of text that does not contain the search word/string is totally ignored. As very few components of the reading model are involved in scanning, there is scarce meaning building at the clause level at most. As such, considering the contextual needs, scanning was not found to be a major skill to test, but can be useful to the students as a time-saving strategy in locating specific information in a text in a short time.

Another practice in reading for orientation is search reading. In search reading, the reader has a predetermined topic in her mind and she wants to locate that information quickly and selectively. The reader may make use of her knowledge of text structure to help her search the pre-specified texts/topics. Once that is done, the reading mode changes into careful reading in order to establish the meaning of a sentence, a paragraph or more. The aim here is to find relevant information quickly in order to answer a question. If the information is found within a single sentence, it is called *search reading local*; if more than a sentence is required to obtain the necessary information it is *search reading global*.

In a testing situation, the reader searches for keywords indicated in the test item or for words in the same semantic field. When she locates the information, she starts to read carefully to establish the propositional meaning at sentence level. In some texts, she may have to integrate information across sentences and make inferences. Testing search reading strategies would yield information about the reader's ability in handling long texts for the purpose of locating and identifying the information she needs.

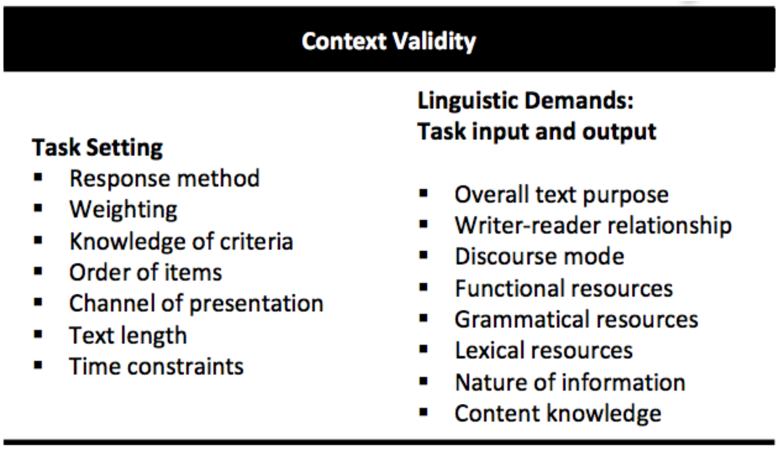
The major reading activity that emerged from the needs analysis study was **reading for information and argument**. This type of reading involves reading as preparation for a course, or an exam, or as an initial step to fulfil an assignment, such as conducting a project or writing a paper, or report. Comprehension of whole text(s), evaluation, synthesis and analysis are taxonomic skills expected in this type of activity. The reader should be able to understand main ideas and details in lengthy texts, understand implicit ideas and writer's stance. Incorporating information from a number of reading sources is also an important academic activity.

Reading for information and argument requires careful reading mostly at the global and rarely at the local level. In this type of reading, the reader comprehends the complete meaning within the text. If there is an unknown vocabulary item, the reader may try to decode the meaning through careful reading at sentence level (local). Hence, in reading for information and argument, almost all of the processes mentioned in the reading model are activated.

4.2.3 Test specifications. After establishing the theoretical basis for reading and elaborating on a new reading model, test specifications were composed (APPENDIX E).

Context validity in the socio-cognitive framework, first discussed in Weir (2005) and later modified in Khalifa and Weir (2009) requires the investigation of parameters under two heading: **Task Setting** and **Task Input and Output** (previously discussed in Chapter 2). (Table 16). In composing the specifications for the test, those parameters guided the design and content of the document.

Table 16 Context validity adapted from Khalifa and Weir (2009)



Context Validity	
Task Setting	**Linguistic Demands: Task input and output**
Response method	Overall text purpose
Weighting	Writer-reader relationship
Knowledge of criteria	Discourse mode
Order of items	Functional resources
Channel of presentation	Grammatical resources
Text length	Lexical resources
Time constraints	Nature of information
	Content knowledge

4.2.3.1 Task setting.

4.2.3.1.1 Response method. Choice of test response method is closely related to the aspect of ability that is being assessed; in other words, the test developer needs to decide whether a specific chosen response method can elicit a specific behavior, or cover certain content (Alderson et al., 1995; Brindley, 2001; Fulcher & Davidson, 2007; Anthony Green, 2014; Haladyna & Rodriguez, 2004). Therefore, it is necessary to identify what type of test task formats can be used to test certain reading types, and what cognitive processes they may activate (Khalifa & Weir, 2009). However, sometimes administrative decisions may prevail. Due to reasons such as objectivity of scoring (test reliability) and lack of resources (scorers and time), it was decided by the school administrators that the reading test would contain selected response type items to a great extent. Among the range of formats in selected response items multiple-choice and multiple matching items were used in the reading test (See Table 10 and Table 20 for the distribution of item types in Reading Test V1 and V2).

Multiple choice format has been widely used in large scale testing due to obvious advantages it offers in grading, i.e. objective marking of items; moreover, the time management of the administration, and grading of the exam becomes easier (Haladyna & Rodriguez, 2004; Khalifa & Weir, 2009). Using multiple choice items in testing high-level processes such as inferencing is much easier due to the control it provides compared to open-ended items (Khalifa & Weir, 2009). Some other advantages of the multiple-choice format frequently mentioned in the related literature can be summarized as:

- it is familiar to the test takers and they know exactly what is expected of them and how to respond,
- using multiple choice questions increases the reliability of the test as a large number of questions can be answered in a limited amount of time, compared to, for example, open ended questions,
- multiple choice questions can be pre-tested easily; hence, they make it easier to set the difficulty level of the test (Weir, 1983)
- multiple choice questions are marked objectively, and therefore, using them increases scoring validity,
- the reading score is not contaminated by other skills (such as writing, as in an open-ended question).

The shortcomings of using the multiple choice format are well-known and are frequently voiced in the literature. Test takers usually approach multiple-choice questions with the intention of solving a problem, for instance, rather than comprehending a text/task. What is more, while selecting options test takers may use test response strategies that may not be relevant in the non-testing context (Rupp, Ferne, & Choi, 2006).

Still, multiple-choice type items are believed to activate processes that resemble the natural processes used during careful and expeditious reading (Khalifa & Weir, 2009), and hence, were used extensively in our reading test. Out of 30 reading items, 16 were prepared in the multiple choice format. They were used to test mainly global careful and global local reading.

Matching is also a selected response item type and it is also scored objectively. This item type was chosen for the testing of mainly careful expeditious and sometimes careful global reading. The matching items asked the test takers to match headings with paragraphs. Out of 30 items, there were 6 matching items.

Short answer items were used to assess the search reading skills of test takers. Short answer items require test takers to write down their answers on the test paper, in a phrase or at most in a sentence. Short answer format was chosen for search reading because it better reflected real life tasks in reading. This skill also "seem[s] to be more testable by short answer questions than multiple choice, the latter involving more spotting and matching of material in the text with the options" (Weir, 1983, p. 339).

In search reading, the readers are expected to read expeditiously to identify the location of the topic they wish to read about. Then, they would read carefully to extract the necessary information. Testing this skill using any of the selected response item types can lead the test taker to scan the text for keywords they find in the options. This would defy the purpose of testing search reading. Therefore, it has been decided that having the test takers produce the answer themselves through the processes they would use in real-life would be a better solution.

One downside of using short answer item type is about grading; inevitably, the papers need to be clerically marked, which brings about some problems such as subjective grading and human error. In order to avoid some of the problems associated with subjective grading, the questions were constructed in such a way that the test takers wouldn't need to make any change in the syntax of the expected answer; they only needed to copy the phrase(s) or clause that they think answers the question on to their answer sheet. This would, hopefully, prevent test takers from grammar pitfalls such as using wrong form of words. As for human error, each paper was scored by two graders to minimize human error.

4.2.3.1.2 Weighting. Weighting is about the points allocated to each item in a test. In case where a certain task is believed to be more important than the others, and it puts a higher cognitive load on the test taker, then that kind of a task may be awarded a different score. If, in a test, there is differential weighting, it needs to be revealed to the test takers beforehand so that they may decide on how to use their time in the goal setting phase (Khalifa & Weir, 2009). The weighting used in our reading test was the same for all types of questions; i.e. all questions were given 1 point.

4.2.3.1.3 Knowledge of criteria. Khalifa and Weir (2009) claim that test takers need to know beforehand the criteria for the judgement of their reading skills. In selected response item type, it is only the judgement of whether they have marked the correct option on the answer sheet. However, in constructed response item type, the test takers should be informed whether they will, for example, lose any points for misspelling a word, or using wrong punctuation.

The selected response item type is included in the careful reading section of the exam, where the test takers are asked to mark the option that they think answers the question correctly. In the search reading part, the test taker is expected to write their responses in the space provided. The expected answers are usually a word, a phrase or a short clause. Small mistakes that do not cause a misunderstanding on the part of the reader are ignored. Those may be spelling mistakes which are usually considered copying mistakes, mistakes in punctuation, missing words – other than content words – and some language errors that do not inhibit the reader from understanding the given response.

4.2.3.1.4 Order of items. According to the processing model of reading proposed in this study, in real-life careful global reading, after processing the sentence, the readers start building a mental model of the text. As they keep reading and receive new information, they integrate it into the representation they have created so far. This modelling suggests that reading is an additive process. It is, therefore, reasonable to present the questions in the same order of the text, consistent with the comprehension process in careful reading (i.e., questions related to the first paragraph of the text are placed before questions related to the second paragraph).

In expeditious reading, items may not be given in order mirroring the text because, the readers sample the text randomly to get its gist. Both top-down and bottom-up processing takes place in expeditious reading, and after establishing a rough macro structure of the text, the readers usually switch to careful reading of the parts that they are interested in. This suggests that items that require expeditious global reading may come first in a reading task, followed by careful global reading items.
In search reading, again, reading is not a linear process: the reader searches for a predetermined topic in any direction until she finds it. Afterwards, she reads carefully for information.

In the reading test, since the majority of the questions require careful global reading, the items are ordered according to the order of the information in the text. In search reading, too, the same format is followed, as the search reading text is much longer than careful reading texts (3000 words vs 1000 words).

4.2.3.1.5 Channel of presentation. It has been suggested that the types of non-verbal information presented in a text, and the associations between non-verbal and verbal pieces of information have an effect on the reader; some readers may benefit from having both verbal and non-verbal information such as diagrams, pictures, maps. It helps to reactivate the information previously read if working memory capacity is limited (Hegarty & Just, 1989; Holliday, Brunner, & Donais, 2018; Khalifa & Weir, 2009).

Despite these advantages, information in our reading test is presented only verbally in the reading test, which is mainly due to habit rather than an informed choice. The testing committee has been informed of the possible positive impacts of presenting non-verbal information on the test-takers and this may be considered in the future. .

4.2.3.1.6 Text length. The decision regarding text length needs to be based on the operations that the test intends to measure (Khalifa and Weir, 2009). Whereas a long text seems to be more appropriate for testing global expeditious and search reading abilities, a shorter text could more easily be processed for reading details.

The reading requirements in the target language use domain also affect the decisions on text length; for example, if an undergraduate student is generally expected to do 10 – 15 pages of reading to prepare for a course each week, it would be logical to test her expeditious reading skill using a text similar in length. However, it is usually not possible to replicate the real life tasks exactly due to a number of practical constraints (e.g. time); therefore, when such decisions are taken resources should also be considered.

As a generalization, one may assume that in academia reading for orientation requires processing of long texts to understand the gist, and reading for information and argument requires reading carefully to understand both the main and supporting ideas in shorter passages. The needs analysis report revealed that close to 90% of the freshman students were required to process texts from a few pages to whole chapters (25 – 30 pages long) weekly as background reading for their courses. A recurring complaint of faculty members teaching freshman courses was that the students generally neglected to read the assigned texts. Among other reasons, lack of knowledge of reading strategies could be a factor to explain this attitude. The longer the texts the more difficult it becomes for the reader to process due to the linguistic and content knowledge required to process them (Skehan, 1998).

Similar complaints were encountered regarding students' abilities in reading critically. Though much shorter texts (one page at most, usually shorter passages from newspaper or journal articles) were used during the classes for critical reading, the students had difficulty in dealing with them.

Seeking a balance between future needs, resources and the cognitive load that reading imposes on test takers, text lengths were decided as follows: For careful reading, texts with 700-1000 words, with at least 7 paragraphs, were found to be appropriate to include the necessary number of items that are required for each reading task (one question per paragraph). This approach (one question per paragraph) is similar to that of some international language tests (e.g. TOEFL:

https://www.ets.org/Media/Tests/TOEFL/pdf/ SampleQuestions.pdf). When the assessed reading ability is careful reading, test takers are expected to read for detail (shorter text – due to time constraints), however, to assess expeditious reading skill, longer texts are needed. Hence, for search reading, 2500 – 3000 words were found to be appropriate to test expeditious and search reading abilities, within the time period reserved for the test. A similar approach to testing search reading was observed in other EFL contexts (see, for example, http://www.yadyok.boun.edu.tr/buept/buept-ornek.htm#search-reading).

4.2.3.1.7 Time constraints. The time given to answer the reading items needs to be allocated carefully so as to obtain reliable results. There are multiple studies on reading speed in the literature, some of which present conflicting results (Table 17).

Table 17 Reading speeds in English

Reading type	**L1 / L2**	**Words per minute**	**Researcher(s)**
Careful R.	L1	250	Nation (2009)
Skimming	L1	500	
Not specified	L2	86.5	Haynes and Carr (1990, in Khalifa and Weir, 2009)
Reading for comprehension	L2	63.5	
Not specified	L1	254	
Not specified	L2	200	Nuttal (1996)
Expeditious R	L1	800	
Not specified	L2	under 100	Jensen (1986)
Expeditious R	L1	800	Heaney (2009)

Considering the recommendations from the literature and the reading speeds of the instructional materials at the DBE, for the careful reading tasks of the test, approximately 60 wpm, for the expeditious reading task a reading speed of 100 wpm was calculated while designing the tasks. While the time allocated for each reading type may seem too much, the time required to read and to answer the questions (including writing short answers) were also included in this timing.

4.2.3.2 Linguistic demands: Task input and output. Khalifa and Weir (2009) argue that the linguistics demands in a test need to simulate those of real-life tasks as closely as possible. Only then, decisions that are based on score interpretations can be justified. For this reason, in making decisions regarding linguistic aspects of input task, several factors were considered such as the demands of the target language context, that is, the first year reading requirements in academic programs, the current test taker profile, and practical issues regarding the administration of the test. By developing a test containing similar linguistic features as texts that the students read in real life, we try ensure that the scores obtained from the test can be generalized; that is, the scores can have predictive value for contexts other than the testing situation.

4.2.3.2.1 Overall text purpose. Overall text purpose (i.e. text function) is closely related to six factors according to Jakobson (1960). He claims that the following factors determine the function of text: addresser, addressee, context, message, contact (between addressor and addressee) and code. Each factor corresponds with a text function respectively: referential, expressive, conative, poetic, phatic, and reflexive. Jakobson (1960) demonstrates the connection between these two sets of variables as follows (Figure 12).

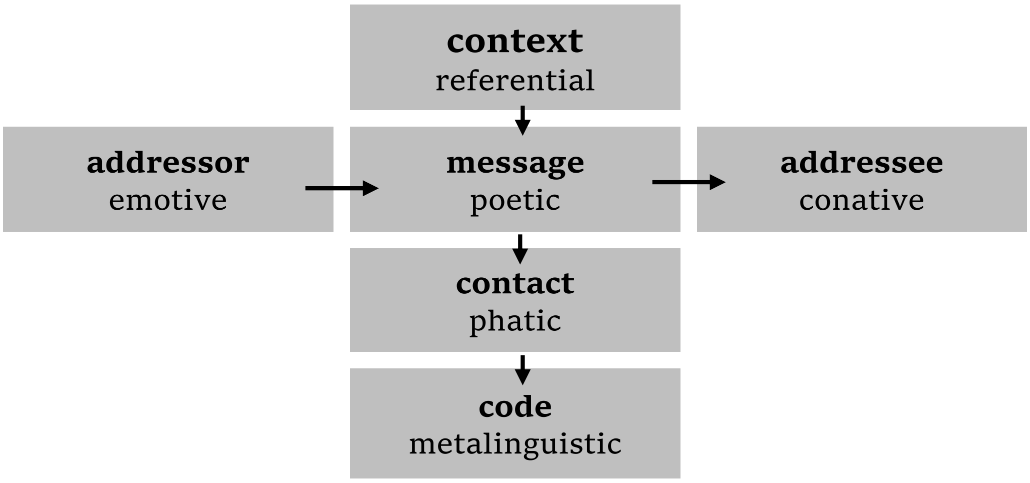


Figure 12 Relation between factors and functions

- The Referential Function is related to content and it intends to inform the reader about a situation, object or mental state.

- The Emotive Function is related to the addresser (speaker) and intends to convey feelings or emotions.
- The Conative Function engages the addressee (receiver) and intends to persuade or convince.
- The Poetic Function focuses on the message and intends to entertain or please.
- The Phatic Function is the use of language for interaction and is related to the contact factor. It intends to keep in touch.
- The Metalinguistic Function is the use of language (code) to discuss of describe language itself. It is language about language.

The referential function is encountered in academic writing as expository text (e.g. definitions, academic essays, book reviews and commentaries) (Vahapassi, 1982), and it is the most common function in academic writing (Khalifa and Weir, 2009). Another essential function of academic writing is argumentation, in other words, the conative function. This function is mostly found in argumentative/persuasive writing (e.g. editorial, critical essay/article) (Vahapassi, 1982). Khalifa and Weir (2009) claim that emotive and phatic functions of written communication can be seen, to a lesser degree, in text messages, blogs, etc.

The texts chosen for the reading test are primarily referential in purpose, that is, they aim to inform the reader on a subject, and sometimes conative, presenting an argument over a subject. These two functions provide ample opportunity to the item writers in creating items in accordance with the purposes of reading such as reading to find the main points in a text with a referential purpose, or following the main arguments in a conative text, in careful global reading.

4.2.3.2.2 Writer-reader relationship. There is reference to the centrality of the reader's capacity or role while creating or choosing a text for a specific purpose (Grabe & Kaplan, 1996; Hyland, 2002). If only the intended reader's capacity is well estimated can the writer decide on the content of the text: with an expert audience in mind, the writer may omit some information relying on the audience's background knowledge or capacity to infer. This decision may also influence the choice of vocabulary and the complexity of grammatical structures (Khalifa and Weir, 2009). As such, a text by itself is not a determining factor; the reader's characteristics have an impact on how and how much of a text is comprehended.

As the audience of the texts in this reading test may come from many different backgrounds, the age-range was taken as a deciding factor in choosing the texts. In addition, care was taken to avoid culturally sensitive material. All texts were aimed at non-specialized, general reader.

4.2.3.2.3 Discourse mode. There are numerous studies that have investigated the relation between comprehension and text organization. Carrell (1987), for example, claimed that when readers are given texts with familiar content and rhetorical form, it influences their comprehension positively. (Urquhart & Weir, 1998).

Meyer and Freedle (1984) studied how the different discourse types resulted in differences when processing the text. They focused on the rhetorical organization of expository texts having collection, description, comparison, problem/solution, and causation structures. Their finding revealed that description and collection structures are generally not as organized as problem solution, causation and comparison structures and that, causation and comparison structures are better recalled than the others.

The CEFR (Council of Europe, 2009) provides a detailed list of text sources for the four contexts of language use: personal, public, occupational and educational. The two domains that were found to be relevant for text selection in the context of this study were educational and public domains. For the educational domain, the sources of texts were specified as textbooks, reference books, journal articles, abstracts, etc. and for the public domain, they were public announcements, labels, notices, regulations, programmes, etc. However, the CEFR does not elucidate how to determine the type of texts appropriate for a specific level in the CEFR. In other sources (A. D. Cohen & Upton, 2006; Khalifa & Weir, 2009), there is reference to different rhetorical tasks encountered in academic texts such as exposition, argumentation and narrative.

Taking account of the recommendations in the literature (Bruce, 2008; Hyland, 2008; Hyon, 1998; Swales, 2004) and the types of texts used in academic programs, the genres for the reading test were decided as academic journals, newspaper and magazine articles, extracts from books, and other informational sources (blogs, internet articles, etc.), and the rhetorical tasks were chosen as descriptive, narrative, expository and argumentative.

4.2.3.2.4 Functional resources. Functions of language were defined in the CEFR (2009) as micro and macro functions: the former refers to short utterances, such as the turns of speakers in a conversation whereas the latter, macro functions, refers to extended text, either spoken or written, with categories such as:

- description
- narration
- commentary
- exposition
- exegesis
- explanation
- demonstration
- instruction
- argumentation
- persuasion (Council of Europe, 2009, p. 126).

Khalifa and Weir (2009) mention some basic functions specific to any CEFR level. At the B2 level, which is defined as an *independent user level*, language users can carry out communicative tasks in various domains (educational, vocational, personal). Van Ek and Trim (2001) provide an extensive list of functions that language users can carry out at B2 level (e.g. describing, narrating, giving opinions, synthesizing, evaluating, critiquing). Khalifa and Weir (2009) claim that it is mainly the lexical and grammatical resources that are needed to express a specific function that are significant in determining the level of the language user. In the reading test, functions commonly attributed to B2 level as well as some low frequency lexical items and complex/compound grammatical structures were included.

4.2.3.2.5 Grammatical resources. Grammatical resources needed at B2 level are closely related to the functions of language identified for this level. As a general rule, the test takers are expected to understand all the main tense forms and grammatical patterns. Complex, and more frequently, compound sentences were present in the texts related to the tasks in Reading Test V2.

As one measure of grammatical complexity, the Flesch Kincaid grade level was computed for all texts (see Table 2 and Table 7 for the Flesch Kincaid grade level of the texts used in Reading Test V1 and V2). As a general rule, it was decided that grade levels 8 -14 were acceptable for use as a reading text at freshman level. As another measure, the number of words in a sentence and number of syllables in a word were also taken as an estimate of text complexity (Khalifa and Weir, 2009). Referencing (comprehending synonymy, adverbials, etc.), complex verb forms, and inferencing were found to be common in texts that are used in Cambridge ESOL exams at B2 and C levels (Khalifa and Weir, 2009); therefore, items inferencing skills were used in our tests.

4.2.3.2.6 Lexical resources. In identifying the expected lexical complexity of texts at each level, the CEFR document provides little help. There are only some general guidelines, without much information on the breadth and depth of vocabulary that might be needed at different proficiency levels (Khalifa and Weir, 2009). Nonetheless, the item writers need guidance in deciding the range of vocabulary that is needed to comprehend freshman level reading texts. One informing source is Laufer (1992), who posits that in order to read satisfactorily in L2, knowledge of over 5000 word families is required. A word family comprises

> the base form of a word and its inflected forms (third person -s, -ed, -ing, plural-s, possessive -s, comparative -er and superlative -est) plus derived forms made from certain uses of the following affixes (-able, -er, -ish, -less, -ly, -ness, -tho -y, non-, un-, -aI, -alion, -ess, -jul, -ism, -ist, -ity, -ize, -menl, in-) (Hirsh & Nation, 1992, p. 692)

Nation (1990) conducted a study in the academic context and found that knowledge of 2800-3000 word families would enable L2 learners to comprehend academic texts. There is also the question of how to choose the 2800-3000 word families, for the non-native learners in the undergraduate context.

So far, several vocabulary lists that define lexical range according to their frequencies have been widely used. They are based on corpus studies, i.e. frequency analyses of words.

The first general list of most frequent words in English, named, the General Service List (GSL), was created in 1940s by Michael West. It contained the most frequently used words in English (P. Nation & Waring, 1997; Schmitt & McCarthy, 1997). Much later, in 2000, Coxhead (2000) developed an academic word list, *An Academic Word List (AWL)*, which was compiled from a corpus of 3.5 million words of written academic text. The AWL contains 570 word families covering approximately 10.0% of the total words in academic texts.

In 2013, Browne et al (2013b) worked on two different lists: *The New General Service List (NGSL)* and *The New Academic Word List (NAWL)*. While working on the NGSL, the researchers claim that they followed the same steps that, West and his colleagues did when creating the GSL, but used objective scientific measure, and pedagogical insights to create a list of 2800 high frequency words with the following goals:

1. to update and greatly expand the size of the corpus used (273 million words compared to the 2.5-million-word corpus behind the original GSL), with the hope of increasing the generalizability and validity of the list
2. to create a list of the most important high-frequency words useful for second language learners of English, ones which gives the highest possible coverage of English texts with the fewest words possible.
3. to make a NGSL that is based on a clearer definition of what constitutes a word
4. to be a starting point for discussion among interested scholars and teachers around the world, with the goal of updating and revising the list based on this input (in much the same way that West did with the original GSL)

The second list that Browne et al. (2013a) created, the NAWL, contains around 960 most frequently used words in academic texts. The list was based on an academic corpus which comprised academic journals, non-fiction, student essays, & academic discourse from the Cambridge English Corpus (CEC), hundreds of top-selling academic textbooks, Michigan Corpus of Academic Spoken and British Academic Spoken English texts.

The NAWL consists of 960 academic words, and the NGSL has around 2800 words. When added together, they cover 92% of the words in the 283 million word in the academic corpus, which is a higher coverage than the other lists (Table 18).

Table 18 Comparison of the coverages of vocabulary lists

Corpus	Size	GSL	NGSL	GSL/AWL	NGSL/NAWL
General	273 Million	84%	90%		
Academic	283 Million			87%	92%

The NGSL and the NAWL were taken as benchmarks in analyzing texts for their lexical properties, and the testing committee used 90% cumulative lexical range for the totality of NGSL and NAWL as the lower limit for the reading texts that were going to be used in the reading test. The vocabulary profiles of all texts are given in Table 4 and Table 9.

4.2.3.2.7 Nature of information. The CEFR suggests using abstract concepts/topics in texts only for language users above a certain level of proficiency. It has been reiterated in the literature that abstract information is more difficult to understand than concrete information (Corkill, Bruning, & Glover, 1988). Khalifa and Weir (2009) posit that language learners may find it easier to process concrete information because both verbal and non-verbal systems are evoked during cognitive operations whereas abstract information is limited with the verbal system.

In Reading Test V2 the topics were mostly concrete, e.g. sea animals, a seed preservation facility. There were two other texts which included partly abstract information: one about internet use and the other about personality traits.

4.2.3.2.8 Content knowledge. Khalifa and Weir (2009) claim that when the reader/test taker interacts with the task – sets out to answer the questions on a test – she uses resources from her knowledge base relevant to the question. This interaction between the test taker's knowledge base and the resources demanded by the task reflects the relation between context and cognitive validity.

It has been emphasized that when tests are prepared for a heterogeneous group of students, the texts should be selected with a 'wider appeal' (Weir, 1983). Test takers may have different backgrounds, and this should not be a factor for success or failure in understanding the content of a test. Urquhart and Weir (1998) suggest that some familiarity with the text topic is advisable for test takers to be able to process it by first activating their schemata. Similarly, Alderson (2000b) argues that absence of any background knowledge on the test content will inhibit comprehension; therefore, some familiarity with the content is recommended.

Khalifa and Weir (2009) provide a list of topics that were found unsuitable for the Cambridge ESOL exams. With similar concerns, subjects that have the potential to offend or upset the test takers are listed as follows:

- war, politics, religion,
- national standpoints on subjects such as genocide or minorities
- death, illnesses, natural disasters
- sex, sexism, racism
- drugs, alcohol, gambling.

It is important to provide tasks with texts that are appealing to the test taker population, but not be biased in favor of a group of people. For the Reading TestV2 , texts on four general topics (i.e., sea animals, a seed preservation facility, internet use and personality traits) were selected.

4.2.3.3 Notes on test specs. The test specs document is a summary in technical terms of the aspects of the contextual parameters of the Reading Test presented in this section. The document starts with the overall purpose of the test, and continues with the general specifications of the two parts, careful reading and expeditious reading, of the test.

A skills taxonomy is added to the specifications document to specify how each reading type, i.e. careful reading and expeditious reading, is operationalized in terms of observable skills and strategies. According to Urquhart and Weir's (1998) taxonomy, careful reading is carried out to understand a text (global, top-down and bottom-up processes), to understand lexis (local, bottom-up process) and to understand syntax (local, bottom-up process).

In Reading Test V2, in testing **careful reading**, test-takers' proficiency in understanding a text and understanding lexis were included. *Understanding a text* was operationalized through items that required the test takers to separate main ideas from supporting details, following the arguments in a text, making inferences, and distinguishing generalizations from examples. In order to assess test takers' ability in *understanding lexis*, they were asked to use contextual clues to predict meaning of unknown words, and/or to correctly select the intended meaning of a lexical item. *Understanding syntax* was inevitably tested, though implicitly. In order to understand the propositional meaning at sentence level, the test takers need to be able to make use of their knowledge of syntax and morphology, as given in the model of reading processes (Figure 3).

In testing **expeditious reading skills**, all three components, i.e., skimming, scanning and search reading were tested. Testing of *skimming* at the global level was operationalized through items that required the text takers to establish the macro structure of a text by reading parts of the text that provide clues on the topic and the main idea of the text, such as the introductory and concluding paragraphs, the abstract, etc. *Scanning*, that is expeditious reading at local level, was not tested as a separate skill, but test takers are expected to do this type of reading when search reading. Finally*, search reading* was operationalized through quick reading and scanning of lengthy texts. Test takers are expected to search for the topic given in the question by using strategies such as reading titles and subtitles, reading abstracts, scanning for lexical items from the question or belonging to the same semantic field.

4.2.3.4 Text selection and text mapping. Text selection was carried out individually by committee members in accordance with the test specifications document that specified various parameters for the texts, from the viable sources to discourse types and linguistic features.

The next step was text mapping. The sessions were arranged with the participation of the committee members, members of the SFL administration, and me, representing

Research and Development Unit. After an initial introduction on the purpose of mapping and the course of progression, the aforementioned group carried out text mapping sessions, as part of *a priori* validation procedures (Urquhart & Weir, 1998; Weir et al., 2000).

The aim in text mapping was to specify the points to be covered for careful global reading questions. The timing for the mapping of each text was calculated according to the length of the text. The established reading speed for careful global reading was around 120 wpm, and with a text of 800-900 words, it took about 7 - 8 minutes to process each text. Each participant in the mapping procedure noted down the main message as well as the more important points they could recall after reading. The points that were agreed by at least 80% of the participants were noted down on the text-mapping document given in APPENDIX F. After text mapping, the individual items were prepared and presented to the committee for evaluation.

The operations and the documentation specified in this section followed the *a priori* validation stages suggested in the reading framework by Weir (2005), Urquhart and Weir (1998) and Khalifa and Weir (2009).

This framework proved to be useful in the sense that it can be used as an assessment development framework as well as a validation framework. Moreover, the nexus between the conceptual structure and the operational activities was maintained throughout the test development process which helped to ensure that each of the operations specified for testing reading was meaningful and justified.

4.3 RQ2: What are the Cognitive Processes That Underlie the Construct of the Reading Test?

4.3.1 Retrospective investigation. Reading Test V1 investigated here consisted of four subtests, with either seven or eight items, that tap on various aspects of reading ability during careful or expeditious reading at the global or local level. This first version of the test consisted of 24 careful reading items and 6 expeditious reading items. During the item development phase, the testing committee used a framework for transition from the existing taxonomy of reading skills to the new reading model by

Urquhart and Weir (1998) and Khalifa and Weir (2009). This transition framework included reading skills under five categories: micro level comprehension, macro level comprehension, critical reading, skimming and scanning. The subskills for each category were given under related heading. As the new reading model comprised of a four-cell matrix (see Table 13), the testing committee and I worked together to prepare a transition framework and mapped each skill and subskill in the existing taxonomy of reading on to the new model (Table 19). After adopting the reading model from Khalifa and Weir (2009), the committee used this guiding framework for a period to help them make the transition to a cognitive reading model.

Table 19 A transition framework

Existing Taxonomy of Reading	**New Reading Model**
Micro level comprehension - Understand vocabulary - Understand meaning at sentence level/detail	**Careful Reading—Local** - Textual cohesion - Syntax - Vocabulary
Macro level comprehension - Separate explicitly stated ideas from supporting details - Understand development of an argument - Understand logical organization of a text **Critical reading** - Draw inference, make predictions, understand writer's purpose & attitude	**Careful Reading—Global** - Understand the organization, underlying structure and development of ideas in a text - Obtain and interpret information from text - Draw inference and conclusion, make predictions
Skimming - Read titles, introduction and conclusion paragraphs, glance at words and phrases to identify text type, topic, purpose	**Expeditious Reading—Global**
Scanning - Look for specific words, figures, dates and names	**Expeditious Reading—Local**

The distribution of items of the reading test according to the reading taxonomy is in Table 20.

Table 20 Distribution of item types into the texts

Item types	**Subtest 1**	**Subtest 2**	**Subtest 3**	**Subtest 4**	**Item#**
Skimming	6 items *Matching item*				6
Micro level comprehension	2 items *Multiple choice*	2 items M*ultiple choice*	6 items *Yes/No/Not Given*	1 item M*ultiple choice*	11
Critical reading		3 items *Multiple choice*	1 item *Multiple choice*	4 items *Multiple choice*	8
Macro level comprehension		2 items *Multiple choice*	1 item *Multiple choice*	2 items *Multiple choice*	5

4.3.1.1 Retrospective protocol form. To investigate the cognitive processes that are involved in reading, a retrospective protocol form was used (see APPENDIX A). The protocol form included a list of statements related to the skills and strategies that were part of the reading model upon which the reading test was developed. The form was filled in by the participant students after completing the test, and analyzed quantitatively. The form had four parts:

- **Part A** aimed to collect information about the background of the participants;
- **Part B** aimed to reveal whether the participants read the text before reading the questions and if they did, whether they read slowly and carefully, or expeditiously;
- **Part C**, contained statements referring to cognitive processes and the knowledge sources that a person might use during reading. This part of the protocol form aimed to reveal which processes were activated while searching for an answer for each question in the subtest;

- **Part D** instructed participants to explain where they found the information to answer each question. It aimed to uncover whether the participants read a single sentence to find the answer, or more.

Each of these parts are linked to the reading model and helped in clarifying whether the cognitive processes reported by the test takers correspond to those outlined in the reading model.

4.3.1.1.1 Part A: Demographic information. Demographic information of the participants is summarized in Table 21. As it was believed that the participants' level of proficiency in English would possibly have an effect on the choice of strategies and extent of skill use, the participants were grouped according to their group levels they were assigned to at the DBE, and comparisons were made accordingly. The participants who were assigned to the beginner and elementary level groups at the beginning of the instructional period, and those who were in their second year of language instruction (repeat group) were categorized as Group 1 (GR1), and the participants who were assigned to the intermediate, upper-intermediate and advanced level groups were categorized as Group 2 (GR2).

Table 21 Participant information

	Subtest 1	**Subtest 2**	**Subtest 3**	**Subtest 4**	**Total**
Group 1	52	49	54	46	201
Group 2	49	48	51	51	199
Total	101	97	105	97	400

The number of students who participated from GR1 and GR2 was almost the same (201 and 199, respectively). The average age of the participants was around 19 for all groups. There was an almost equal distribution between male (n=198) and female (n=202) participants. About 50% of all the participants reported that they received English language education at high school. The participants' average mean scores for the reading subtests according to their level groups are given in Table 22.

Accordingly, Subtest 3 was the easiest test for both groups (GR2 M=79.3 and GR1 M=66.7) and Subtest 1 was the most difficult one (GR2 M=46.7 and GR1 M=30.3). (For statistical properties of all subtests and their items, see Table 35.).

Table 22 Mean scores according to level groups

	Subtest 1	**Subtest 2**	**Subtest 3**	**Subtest 4**
Group 1	30.3	38.7	66.7	43
Group 2	46.7	56.3	79.3	69
Average	38.5	47.5	73.5	56

4.3.1.1.2 Part B: Test preview strategies. Part B of the protocol form inquired the previewing strategies used by the participants; that is, whether they read the text before looking at the questions, and if they did, did they read it slowly and carefully, or quickly and selectively. For each test item they were expected to choose from options:

1) read whole or part of the text slowly and carefully (*slowly & carefully*),

2) read whole or part of the text quickly and selectively to get a general idea (*quickly & selectively*), or

3) did not read the text (*no preview*).

4.3.1.1.2.1 Summary results. Table 23 reveals that the most frequently used previewing strategy in Subtest 1 was reading *quickly & selectively* (41%). This means, before looking at the questions participants looked through the text quickly to get an idea about the content of it. Reading *quickly and selectively*, in our reading model, corresponds to the expeditious reading strategy. As the participants did not look at the questions beforehand, it was not possible to establish a relationship between item types and previewing strategy. The second preference in previewing was reading *slowly & carefully* with 32%, and the third was *no preview* with 27%. Reading *slowly & carefully* corresponds to careful reading in the reading model.

Table 23 Preview strategies in the four subtests

		Slowly & carefully	**Quickly & selectively**	**No preview**
Subtest 1	N	247	309	205
	%	32%	41%	27%
Subtest 2	N	187	229	255
	%	28%	34%	38%
Subtest 3	N	236	363	217
	%	29%	44%	27%
Subtest 4	N	206	228	216
	%	32%	35%	33%

In Subtest 2, the most frequently used previewing strategy was no preview with 38%, followed by reading *quickly & selectively* (34 %), and reading *slowly & carefully* (28%). This means, 38% of the participants did not read the text but read the questions before reading the text.

Participants who answered Subtest 3 reported that they read the text *quickly & selectively* (44%) before reading the questions. Those who said they read *slowly & carefully* and who said they did not read the text before they read the questions were similar in percentage (29% and 27%, respectively).

Participants who answered Subtest 4 reported to have used the three strategies in similar proportions: reading *quickly & selectively* with 35%, followed by *no preview* (33%) and reading *slowly & carefully* (32%).

Another analysis regarding test preview strategies was carried out between the responses of those who correctly answered the questions in the subtests and all responses (both correct and incorrect).

Table 24 Proportion of correct responses according to previewing strategies

		Slowly and carefully	**Quickly and selectively**	**No preview**
Subtest 1	All responses	32%	41%	27%
	Correct responses	35%	37%	28%
Subtest 2	All responses	28%	38%	34%
	Correct responses	30%	30%	40%
Subtest 3	All responses	29%	44%	27%
	Correct responses	29%	43%	28%
Subtest 4	All responses	32%	33%	35%
	Correct responses	36%	31%	33%

The comparison of the test preview strategies used by the participants who successfully answered the items with the summative results obtained from all participants' answers reveals slight differences between the two groups (Table 24). In Subtest 1, the majority of the correct responses came from the participants who reported that they read *quickly and selectively (37%)*, followed closely by those who read *slowly and carefully* (35%). Those who did not read the text at all were a smaller proportion compared to the others (28%).

In Subtest 2, the proportion of participants who said they read *slowly and carefully* and *quickly and selectively* were the same (30%). However, the majority of the correct answers came from people who said they did not read the text beforehand (40%),

The answers for Subtest 3 were similar to those of Subtest 1: the majority of the correct responses came from the participants who reported that they read *quickly and selectively* (43%).

Finally, in Subtest 4, the correct responses showed a similar distribution between the three options, with reading *slowly and carefully* having the highest percentage among all (36%).

Bearing in mind the results from all four subtests, there did not seem to be a common tendency among the participants in their use of previewing strategies. All three strategies were reported to have been preferred at least once by all participants.

4.3.1.1.2.2 Previewing strategies by GR1 and GR2 participants. In order to arrive at a mean score to carry out descriptive and inferential statistics, the total number of times the participants reported to have used a certain strategy in each text was divided by the number of participants. The data in each subtest were analyzed using the Mann-Whitney U test to reveal whether the differences between GR1 and GR2 participants in terms of the choice of strategies were significant. Those that were found significant were reported in the relevant sections.

Figure 13 shows the mean scores for the three previewing strategies as reported by GR1 and GR2 participants for Subtest 1.

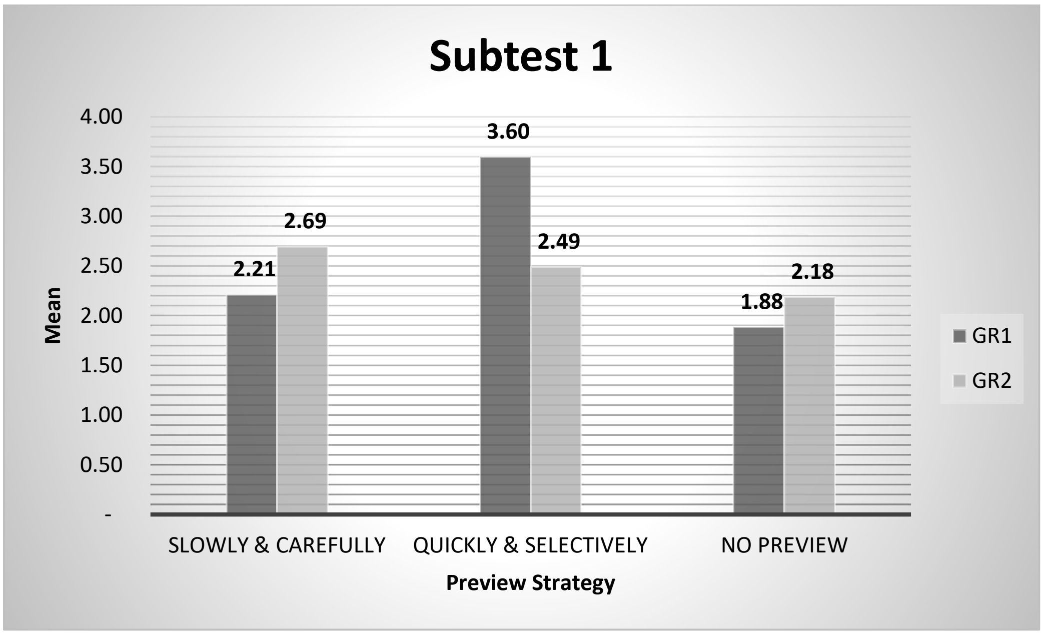


Figure 13 Subtest 1 - Previewing strategies

Accordingly, GR2 participants showed an almost equal distribution of preference for the three strategies; however, the mean score for the strategy *reading whole or part of the text slowly and carefully* was a little higher (M=2.69, SD=3.04) than *reading whole or part of the text quickly and selectively to get a general idea* (M=2.49, SD=2.93) and *no* *preview* (M=2.18, SD=3.24). In terms of GR1 participants, they chose to read *whole or part of the text quickly and selectively to get a general idea* more frequently than GR2 participants did. The mean score of *reading whole or part of the text quickly and selectively to get a general idea* was higher (M=3.60, SD=3.34) than that of the other two strategies. The Mann-Whitney U test revealed that GR1 and GR2 were not significantly different with regard to the use of the strategies *reading whole or part of the text slowly and carefully* and *no preview*; however, GR1 participants used the strategy *reading whole or part of the text quickly and selectively to get a general idea* significantly more than GR2 participants did, U=984, p<.05, *r*=-.20.

Figure 14 shows the mean scores for the three previewing strategies as reported by GR1 and GR2 participants for Subtest 2.

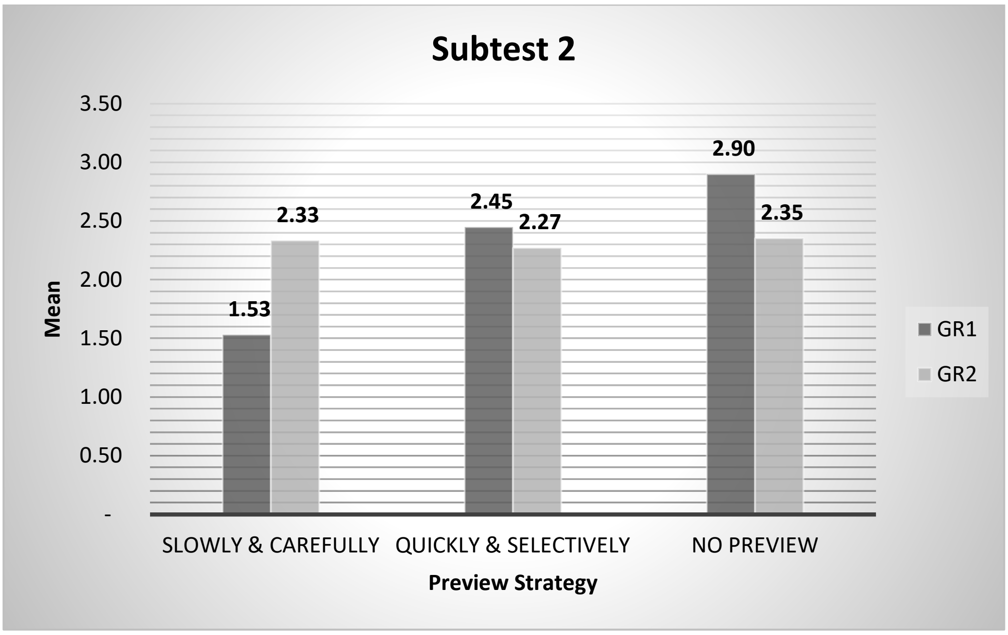


Figure 14 Subtest 2 - Previewing strategies

According to Figure 14, in Subtest 2, GR2 participants' responses had a pattern similar to that of Subtest 1: there was not much difference among the mean scores of the three strategies used by GR2 participants; each strategy was reported to have been used in a similar ratio. However, GR1 participants' use of these strategies showed a difference: the strategy that was used most was *no preview* (M=2.90, SD=3.07), it was followed by *reading whole or part of the text quickly and selectively to get a general idea* (M=2.45, SD=2.68). These two strategies were used more frequently by GR1 participants. The lowest mean score was obtained from the strategy *reading whole or part of the text slowly and carefully* (M=1.53, SD=2.24), again by GR1 participants.

Figure 15 shows the mean scores for the three previewing strategies as reported by GR1 and GR2 participants for Subtest 3. In Subtest 3 analysis, the strategy *reading whole or part of the text quickly and selectively to get a general idea* had the highest mean score for both GR1 and GR2 participants (M=3.69, SD=3.42 and M=3.22, SD=2.59, respectively). For the GR2 participants, *reading whole or part of the text slowly and carefully* and *no preview* were chosen in very similar proportions, around M=2.24. However, for GR1 participants, *reading whole or part of the text slowly and carefully* had a higher mean score (M=2.26, SD=3.10) than *no preview* (M=1.89, SD=3.14).

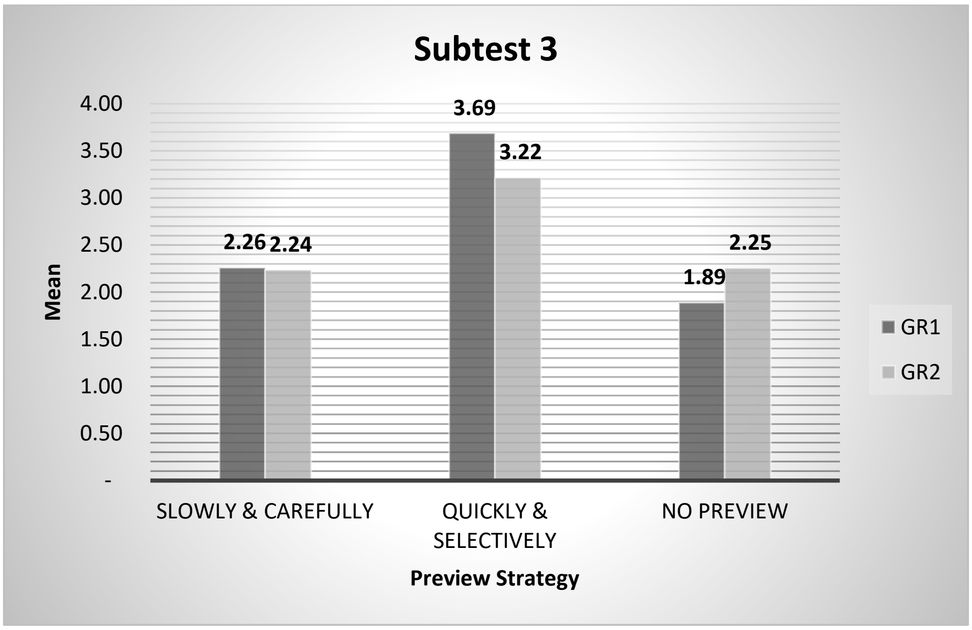


Figure 15 Subtest 3 - Previewing strategies

Analysis of the protocol forms from Subtest 4 (Figure 16)) revealed that GR2 participants had proportionate preference for *reading whole or part of the text slowly and carefully* (M=2.41, SD=2.95) and *no preview* (M=2.53, SD=3.06), but they used the strategy *reading whole or part of the text quickly and selectively to get a general idea* to a lesser extent (M=1.94, SD=2.66). On the contrary, GR1 participants had a strong preference to *read whole or part of the text quickly and selectively to get a general idea* (M=2.80, SD=2.32) compared to the other two strategies. The Mann-Whitney U test revealed that the difference between GR1 and GR2 in the use of the strategies *read whole or part of the text slowly and carefully* and *no preview* were not significant; however, GR1 participants used the strategy *read whole or part of the text quickly and selectively to get a general idea* significantly more than GR2 participants did, U=887.5, p<.05, *r*=-.22.

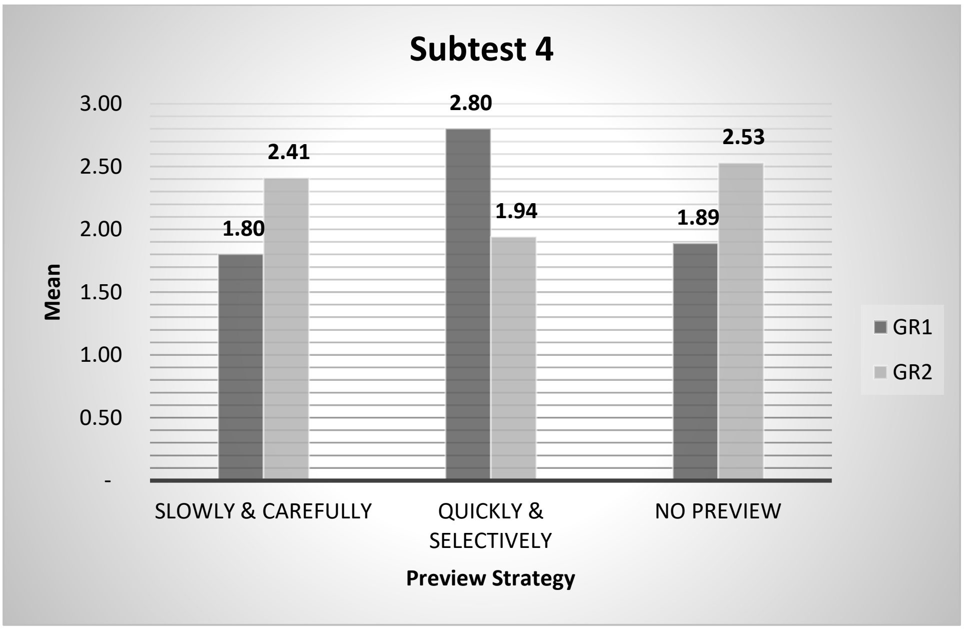


Figure 16 Subtest 4 - Previewing strategies

4.3.1.1.3 Part C: Test response strategies. In Part C of the protocol form, test response strategies were investigated. The test response strategies were also taken from the retrospective protocol form by Weir et al. (2009). The participants reported which skills, strategies and types of knowledge they used to answer each question in the reading test. There were eleven items in this part which were briefly referred to within the visuals as follows:

1. *Scan & match*: match words that appeared in the question with exactly the same words in the text
2. *Search & match similar:* quickly match words that appeared in the question with similar or related words in the text

3. *Writer highlight*: look for parts of the text that the writer indicates to be important
4. *Read key parts*: read key parts of the text such as the introduction and conclusion
5. *Work out word*: work out the meaning of a difficult word in the question
6. *Use vocabulary*: use my knowledge of vocabulary
7. *Use grammar*: use my knowledge of grammar
8. *Read slowly & carefully*: read the text or part of it slowly and carefully
9. *Re-read parts*: read relevant parts of the text again
10. *Use knowledge of organization*: use my knowledge of how texts like this are organized
11. *Connect with prior knowledge*: connect information from the text with knowledge I already have

The data was summarized, first, according to the responses of all participants, and then, according to responses of GR1 and GR2 participants in order to reveal whether there were differences in the patterns of strategy use between students at different proficiency levels. In the analysis, the number of times each strategy was reported to have been used was divided by the number of participants in each group to arrive at mean scores for each strategy.

4.3.1.1.3.1 Summary of test response strategies. Figure 17 shows strategy use in all subtests ordered from the most frequently used to the least. The most popular three strategies in all subtests were S8, *read slowly & carefully*, S9, *re-read parts*, and S2, *search & match similar* (in Subtest 2, instead of S2, S6, *use vocabulary*, was more frequently used).

S8 is a careful reading strategy at the global level and S9, is again a careful reading strategy that might be used at the local or global level. The results revealed that careful reading was one of the most frequently used reader purpose, either at the global or local level. S2 is a scanning strategy which is used to match words that appear in the question with similar or related words in the text. S6, which was frequently used in Subtest 2, refers to the use of vocabulary knowledge while answering the questions.

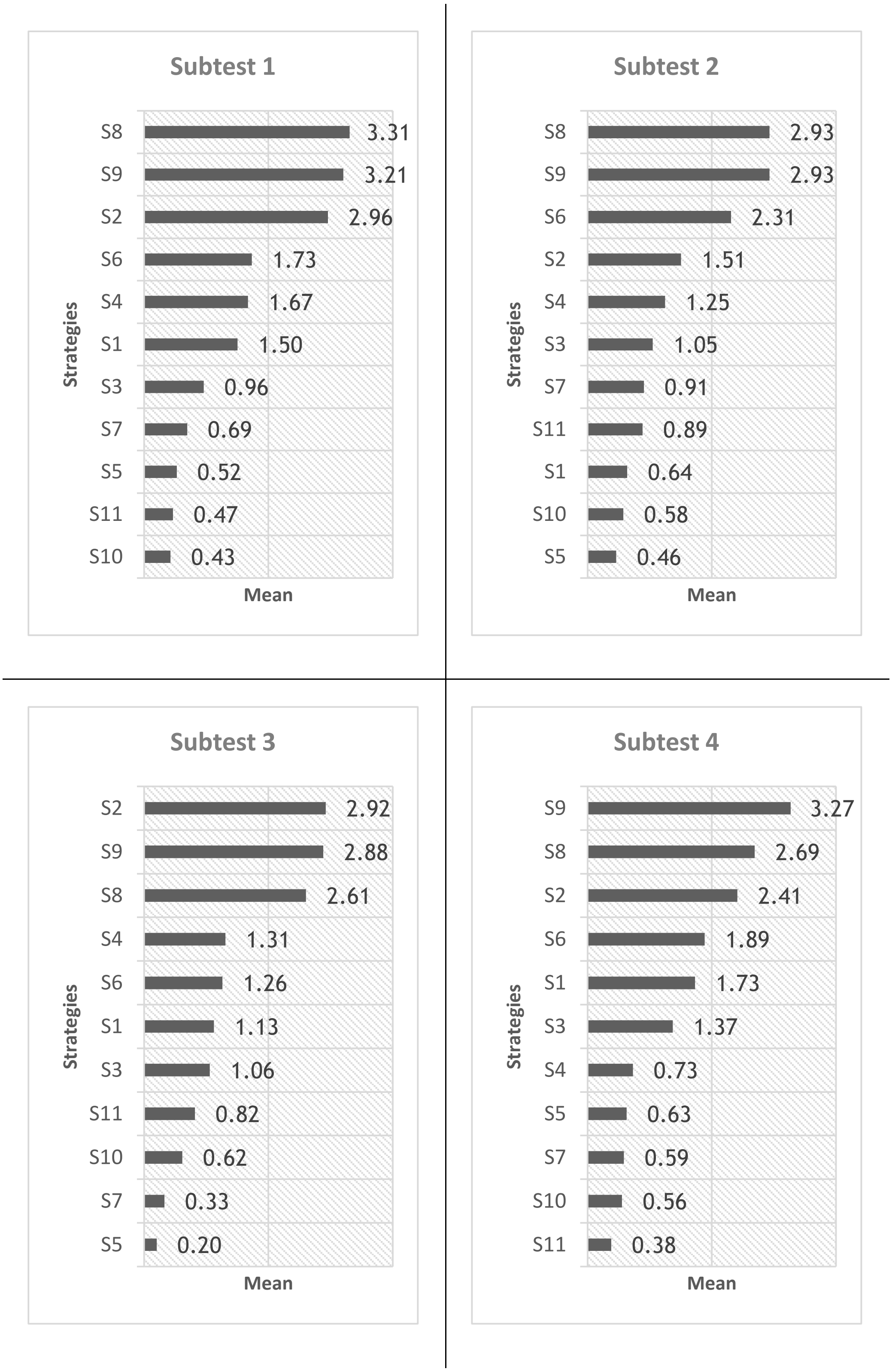


Figure 17 Means of test response strategies - all subtests

The least frequently used strategies were S5, S7, S10 and S11. S5 refers to trying to understand the meaning of a difficult word in the question using knowledge of vocabulary, and S7 refers to using the knowledge of grammar to understand the sentence structures. As these strategies were used infrequently, it was possible to conclude that the questions were easy in terms of their lexical properties and grammatical construction. Another infrequently used strategy was S10, which refers to using knowledge of text organization in finding an answer to a question. Apparently, very few participants used this strategy, probably because in the questions, the respondents were already pointed to the paragraphs where they would find the answer; therefore, they didn't need to search for the location of the answer using knowledge of text organization. S11 refers to using knowledge one already has about the topic while answering the question. Having few respondents choosing this option was a positive outcome since test developers would not want test takers to answer a question using a strategy unrelated to the test construct.

4.3.1.1.3.2 Comparison of GR1 and GR2 participants' responses. In Figure 18, the comparison of the mean scores in strategy use between GR2 and GR1 participants who answered Subtest 1 revealed that both group of participants reported to have used S8, *read slowly & carefully* (GR2 M=3.80, SD=3.00, GR1 M=2.85, SD=2.85), S9, *re-read parts* (GR2 M=3.67, SD=2.79, GR1 M=2.77, SD=2.72) and S2, *search & match similar* (GR2 M=2.96, SD=2.67, GR1 M=2.96, SD=2.54) more than the other strategies. The least frequently used strategy for GR1 participants was S10, *use knowledge of organization* (M=.21, SD=.57), and for the GR2 participants it was *S11, connect with prior knowledge*: (M=.37, SD=1.18).

The strategies that were preferred more frequently by the GR2 participants were: S3, *writer highlight,* S5, *work out word,* S6, *use vocabulary,* S7, *use grammar,* S8, *read slowly and carefully,* S9, *re-read parts,* and S10, *use knowledge of organization*. The strategy S2, *search & match similar* was used with a similar frequency by these two groups; the rest of the strategies were preferred mainly by GR1 participants.

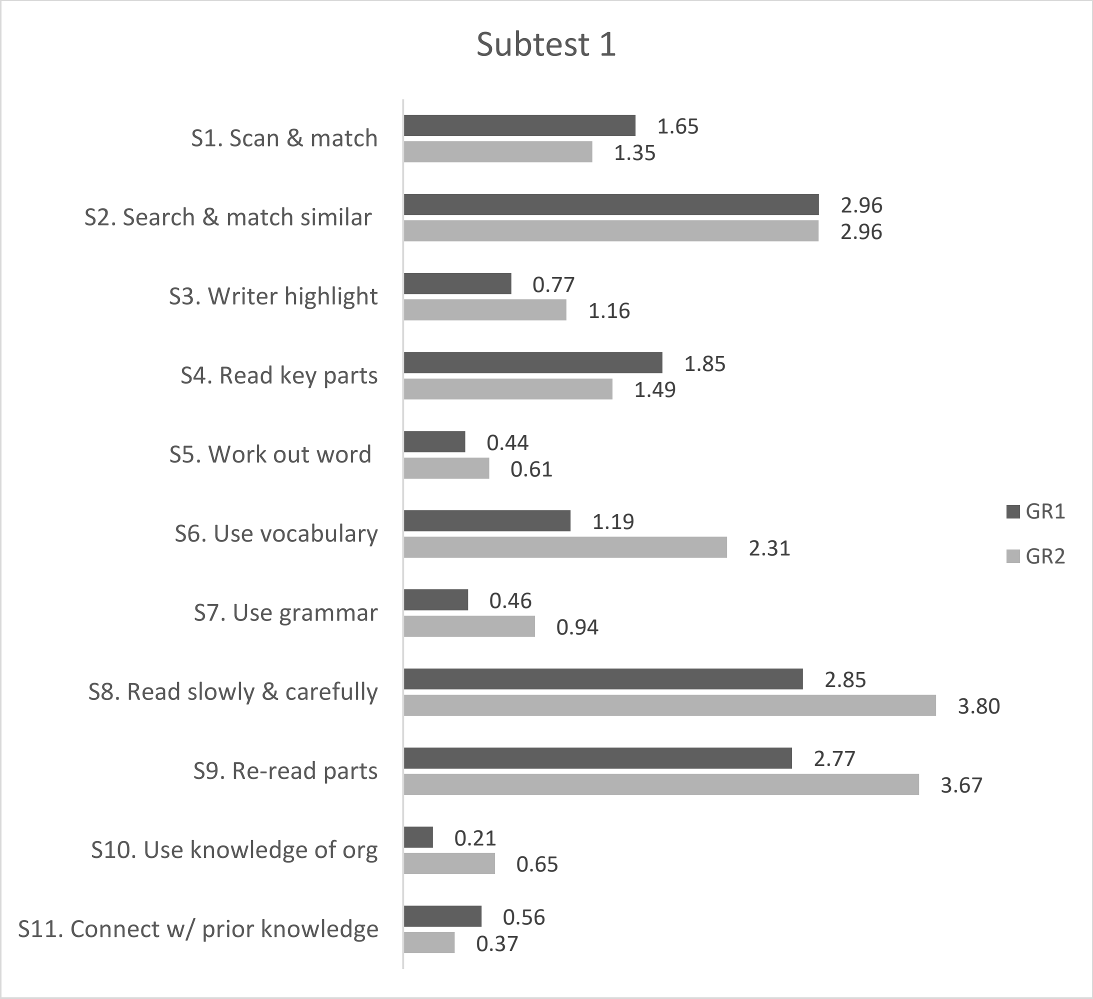


Figure 18 Mean scores of strategies (Subtest 1)

In comparison within groups, the biggest difference was in S6, *use vocabulary* (GR2 M=2.31, SD=2.83; GR1 M=1.19, SD=1.98). The Mann-Whitney U test revealed that the GR2 participants used their knowledge of vocabulary significantly more than GR1 participants did, U=978.5, p<.05, *r*=-.31. There was no other statistically significant difference in strategy use between GR1 and GR2 participants.

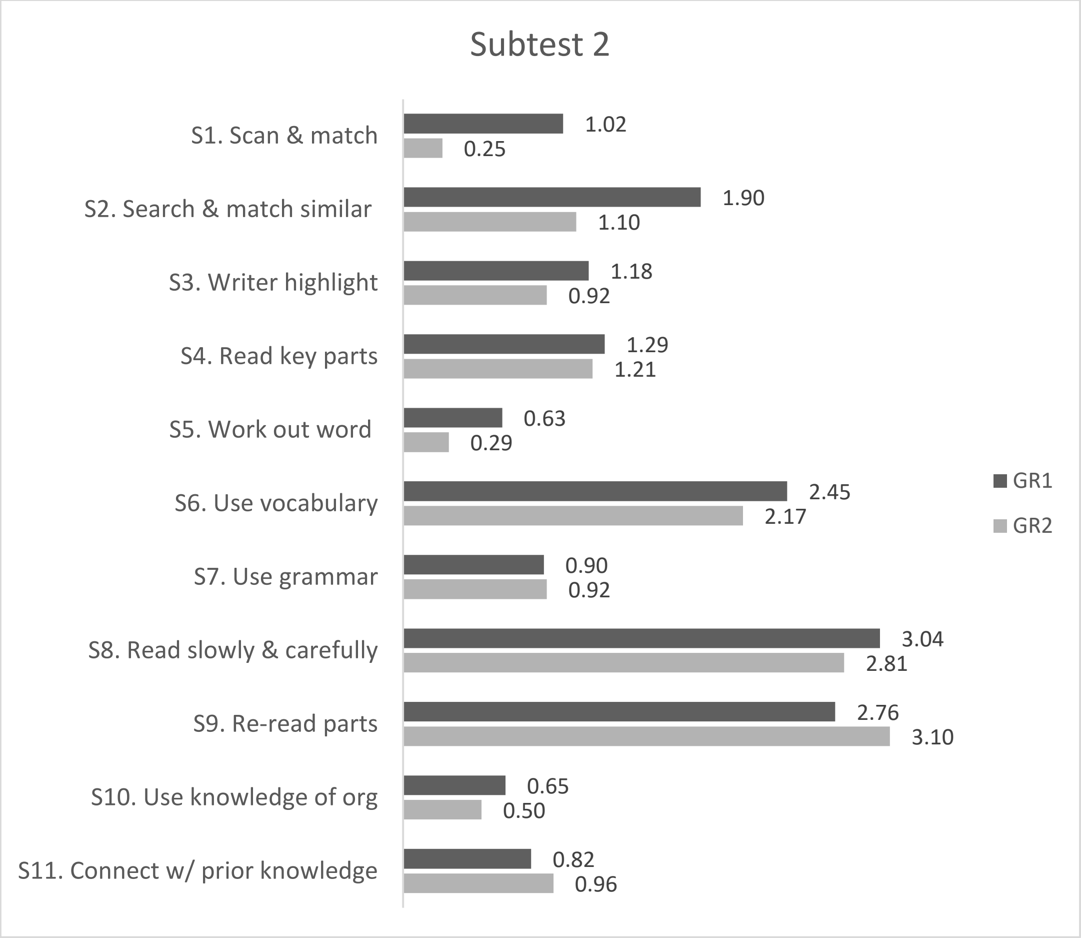


Figure 19 Mean scores of strategies (Subtest 2)

In Subtest 2, the top three strategies that were most frequently used were the same for the GR2 and GR1 participants: S8, *read slowly & carefully* (GR2 M=2.81, SD=2.02; GR1 M=3.04, SD=2.45), S9, *re-read parts* (GR2 M=3.10, SD=2.16; GR1 M=2.76, SD=2.15), and S6, *use vocabulary* (GR2 M=2.17, SD=2.45; GR1 M=2.45, SD=2.64) (Figure 19).

The strategies that were used more frequently by GR2 participants were S8, *read slowly and carefully*, and S11, *connect with prior knowledge*. All the other strategies were used more frequently by GR1 participants except for S7, *use grammar*, which was used almost equally frequently by both participant groups.

The strategy that was used least frequently was S1, *scan & match* for the GR2 participants (M=.29, SD=.53), and S5, *work out word* for GR1 participants (M=.63, SD=1.45).

The Mann-Whitney U test revealed that there were significant differences between GR1 and GR2 participants in the use of strategies S1, *scan and match* and S2, *search and match similar*. GR1 participants' use of S1*, scan and match* was significantly more frequent than the HLP participants' use of that strategy, U=810, p<.005, *r*=.32. In a similar vein, GR1 participants used the strategy S2, *search and match similar* more frequently that the GR2 participants did, U=882.5, p<.03, *r*=.223.

In Subtest 3, the top three strategies that were used most frequently by both groups of participants were S8, *read slowly & carefully* (GR2 M=2.69, SD=2.07; GR1 M=2.50, SD=2.71), S9, *re-read parts* (GR2 M=2.90, SD=2.38; GR1 M=2.85, SD=2.55), andS2, *search & match similar* (GR2 M=2.57, SD=1.85; GR1 M=3.15, SD=2.60) (Figure 20).

The least frequently used strategies were again the same for both groups of participants: S7, *use grammar* (GR2 M=.14, SD=.38 and GR1 M=.54, SD=1.64) and S5*, work out word meaning* (GR2 M=.16, SD=.42 and GR1 M=.26, SD=.52).

In terms of the differences in strategy used between GR1 and GR2 participants: GR2 participants reported to have used two strategies, S9, *re-read parts* (M=2.90, SD=2.38) and S11, *connect with prior knowledge* (M=.88, SD=1.18) more frequently than GR1 participants (M=2.85, SD=2.56 and M=.78, SD=1.59, respectively). All the other strategies were used more frequently by GR1 participants.

In comparison within groups, the biggest difference was in the use of S11, *connect with prior knowledge*, which was found to be significant according to the Mann-Whitney U test. The GR2 participants used their prior knowledge about the topic more frequently than GR1 participants did, U=1054.5, p<.05, *r*=-.23.

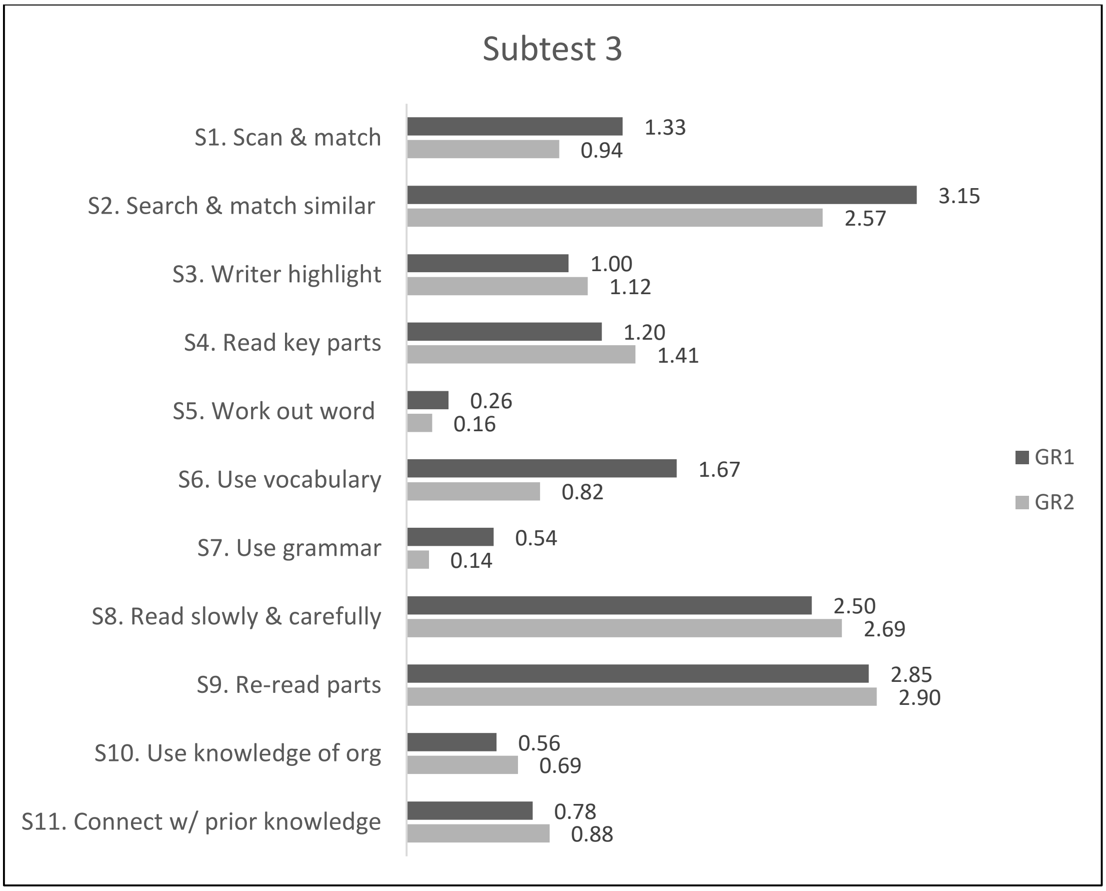


Figure 20 Mean scores of strategies (Subtest 3)

Figure 21 shows GR1 and GR2 participants use of the eleven strategies while responding to Subtest 4. The top three strategies that were used most frequently by the GR2 participants were S9, *re-read parts* (M=3.16, SD=2.23), S8, *read slowly & carefully* (M=2.98, SD=2.45) and S2, *search & match similar* (M=2.18, SD=2.11). For GR1 participants however, the top three most frequent strategies were, S9, *re-read parts* (M=3.39, SD=2.06), S2, *search & match similar* (M=2.67, SD=1.52), and S1, *scan & match* (M=2.39, SD=2.05). The strategy that was used the least was S11, *connect with prior knowledge* for both GR1 and GR2 participants (GR1 M=.3, SD=.70; GR2 M=.45, SD=1.24).

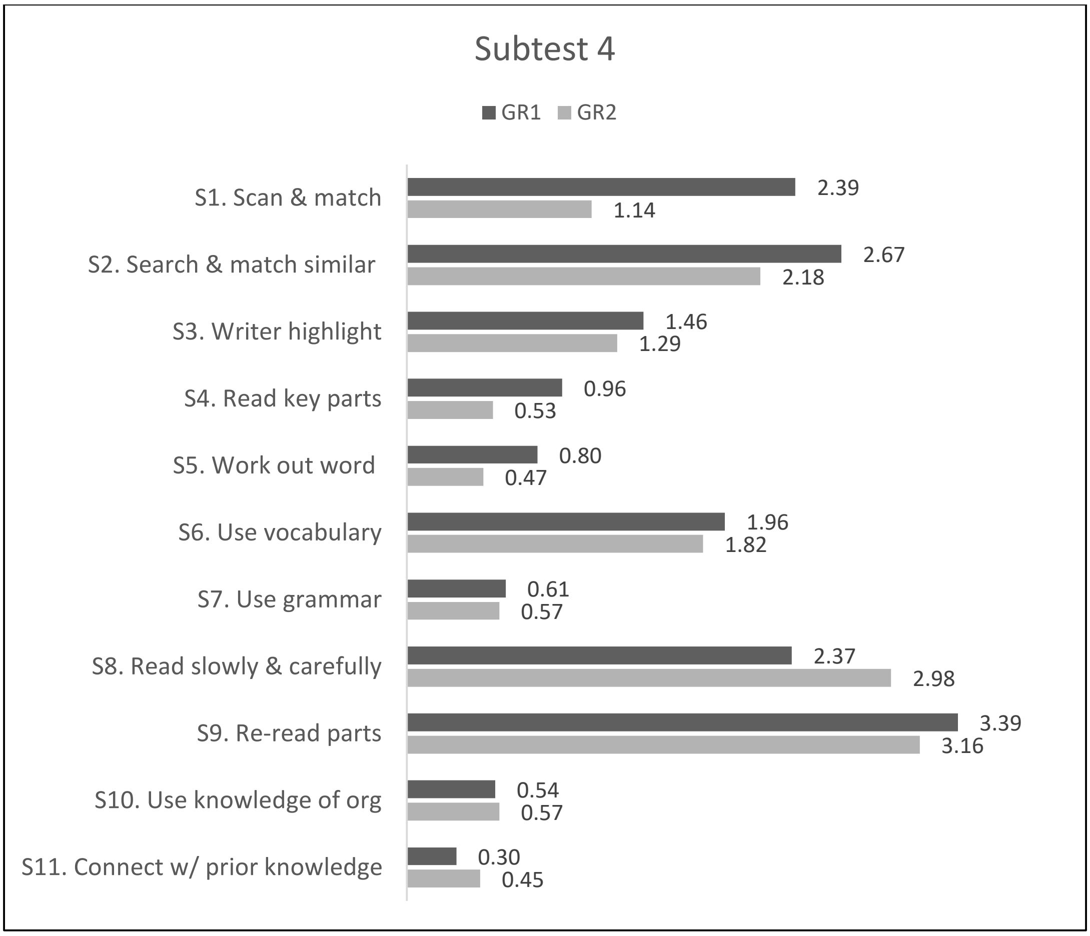
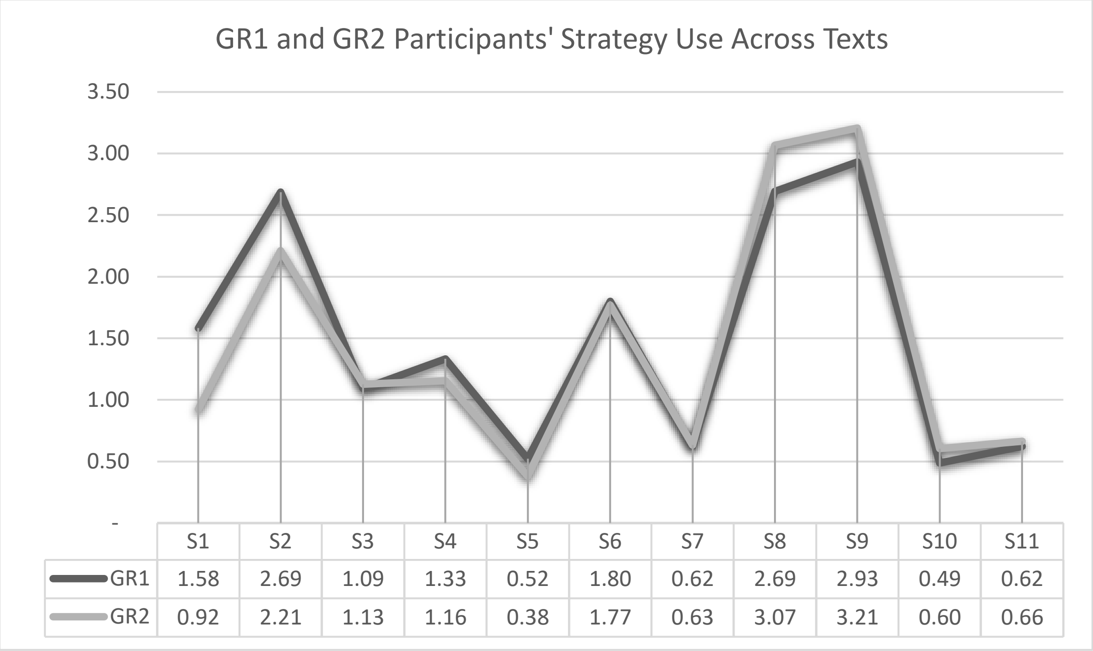


Figure 21 Mean scores of strategies (Subtest 4)

In Subtest 4, only three of the strategies were preferred more frequently by the GR2 participants: S8, *read slowly and carefully*, S10, *use knowledge of organization* and S11, *connect with prior knowledge*. The rest of the strategies were preferred more frequently by GR1 participants and there were some significant differences between the two groups' use of these strategies.

In comparison within groups, the biggest differences between GR1 and GR2 participants were in S1, *scan & match*, S2, *search & match similar* and S4, *read key parts*. The Mann-Whitney U test revealed that GR1 participants matched words that appeared in the question with exactly the same words in the text more frequently than the GR2 participants did, U=692.5, p<.01, *r*=-.53. They also quickly matched words that appeared in the question with similar or related words in the text more frequently than

GR2 participants did, U=883.5, p<.05, *r*=-.31. And finally, the same group of participants S4, *read key parts of the text* such as the introduction and conclusion more frequently than GR2 participants did, U=925.5, p<.05, *r*=-.29.



	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
GR1	1.58	2.69	1.09	1.33	0.52	1.80	0.62	2.69	2.93	0.49	0.62
GR2	0.92	2.21	1.13	1.16	0.38	1.77	0.63	3.07	3.21	0.60	0.66

S1: scan & match
S2: search & match similar
S3: writer highlight
S4: read key parts
S5: work out word
S6: use vocabulary
S7: use grammar
S8: read slowly & carefully
S9: re-read parts
S10: use knowledge of organization
S11: connect w/ prior knowledge

Figure 22 Collated results from four subtests

In order to reveal whether there was a pattern in the strategy preferences of GR1 and GR2 participants, I gathered the data from the four subtests and compared them. According to Figure 22, in all subtests, the frequency of use of different strategies were similar for both groups of participants with some slight differences in a few points.

The data revealed that GR1 mean scores of S1, *scan & match* (M=1.58), S2, *search & match similar* (M=2.69) and S4, *read key parts* (M=1.33) were higher than GR2 mean scores (S1 M=0.92, S2 M=2.21, and S4 M=1.16).

The strategy that was used most by all participants was S9, *re-read parts* (GR1 M=2.93, GR2 M=3.21) followed by S8, *read slowly and carefully* (GR1 M=2.69, GR2 M=3.07) and the strategies that were used the least were S5, *work out word* (GR1 M=.52, GR2 M=.38) and S10, *use knowledge of organization* (GR1 M=.49, GR2 M=.60) (Figure 22).

Figure 23 provides an overall view on the use of strategies across tests, and by both item types, and item objectives. The squares at the intersections of items denote the frequency of use of each strategy: each marked small square corresponds to a 20% usage frequency. For example, four marked square means over 80% use frequency and no marked squares means frequency of use is smaller than 20%.

In terms of item types, in responding to matching items in Subtest 1, participants reported that they mainly used the strategies —from the most frequent to the least — *read slowly & carefully* and *re-read parts* being at the top, *search & match similar*, *scan & match*, *read key parts* and finally, *use vocabulary*. In answering the multiple-choice type test items in Subtest 1, *read slowly & carefully* and *re-read parts* were again the most frequently used strategies. A similar pattern was observed in the responses given to the questions in Subtest 2. In contrast, in Subtest 3, the participants reported that they used *search & match similar* most frequently in answering multiple-choice test items; however, in the rest of the questions, which were Y/N/NG type of questions, *read slowly & carefully* and *re-read parts* were the most popular strategies. Finally, in Subtest 4, similar to Subtest 2, there were only multiple-choice questions and *read slowly & carefully* and *re-read parts* were mostly used. The least used strategies in all four subtests were *use knowledge of organization*, *connect with prior knowledge, work out word* and *use grammar*.

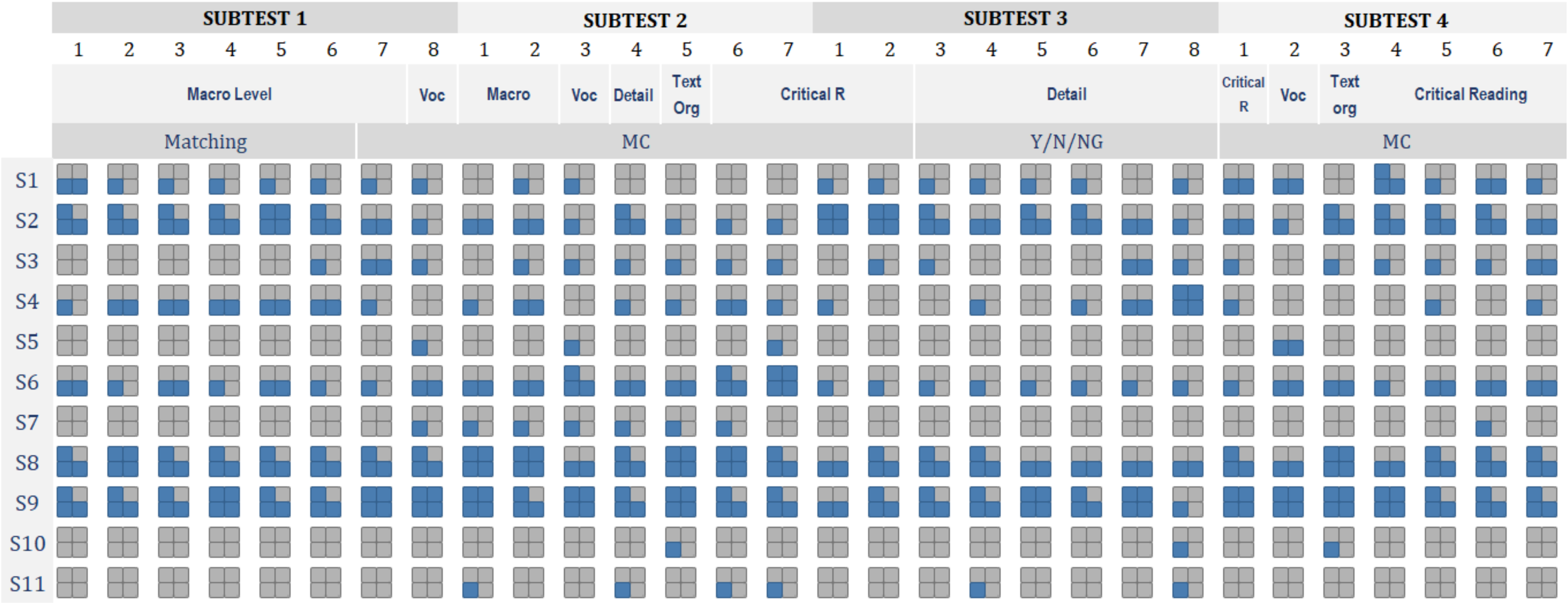


Figure 23 Overall view on the use of test response strategies in the four subtests.

4.3.1.1.4 Part D: Locating information. Part D of the protocol form asked the respondents where they found the answer. The options given in the form were:

L1: *Single sentence*: within a single sentence

L2: *Across sentences:* by putting information together across sentences

L3: *Whole text*: by understanding how information in the whole text fit together

L4: *Without read*: without reading the text

L5: *No answer*: could not answer the question.

4.3.1.1.5 Overview across texts. In Subtest1 and Subtest 2, the participants reported that they found the answer by putting information together across sentences (L2) 49% of the time. In Subtest 3 and Subtest 4, the usage percentage for L2 was even higher, 52% and 56%, respectively.

The second most popular answer to locating information was L3, whole text: the usage percentages were the same for Subtest 1 and Subtest 2 (31%); for Subtest 3 it was 22%, and for Subtest 4, it was 21%.

The results showed that at about 13% – 22% of the time the respondents found the answer within a single sentence, and about 1% - 3% of the time they found the answer without reading the text. The percentage for no answer was between 1% - 6% for all subtests. (Table 25).

Although these findings reveal that the majority of the participants primarily used the strategy L2, *across sentences* in all subtests, it does not tell us whether the use of this strategy was an appropriate choice in answering the questions correctly. Therefore, again, a comparison was made between the strategies used by GR1 and GR2 participants.

Table 25 Results of locating information for the four subtests

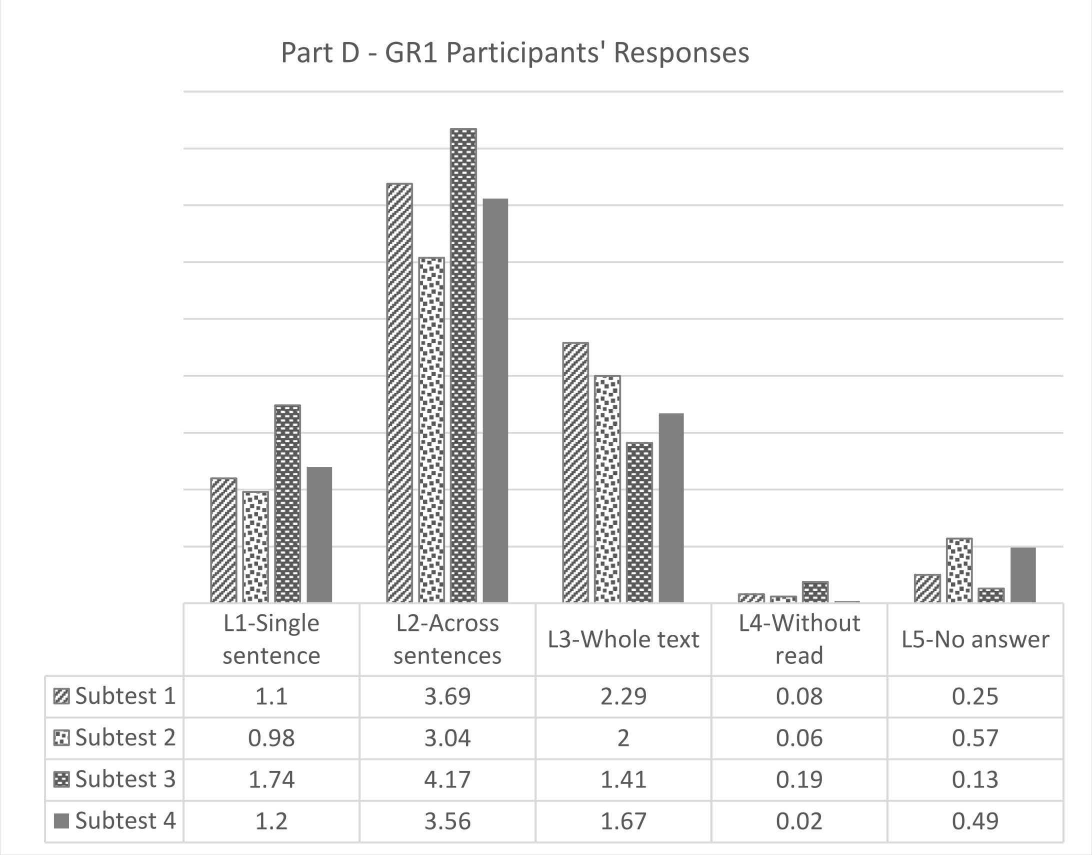
	Subtest1		Subtest 2		Subtest 3		Subtest 4	
	N	%	N	%	N	%	N	%
L1-Single sentence	119	16	83	13	177	22	120	18
L2-Across sentences	372	49	321	49	420	52	368	56
L3-Information fit	233	31	201	31	180	22	139	21
L4-Without read	6	1	10	1	22	3	6	1
L5-No answer	21	3	39	6	8	1	25	4

4.3.1.1.5.1 GR1 participants' responses. According to Figure 24, L2, *across sentences* had the highest mean scores in all four subtests among the five options in Part D of the protocol form. Subtest 3 had the highest mean score (M=4.17, SD=2.17) followed by Subtest 1 (M=3.69, SD=2.55), Subtest 4 (M=3.56, SD=1.62) and finally Subtest 2 (M=3.04, SD=1.68).

The second most frequently chosen option was L3, *whole text*, i.e. the participants found the answer by understanding how the information in the whole text fit together. This option was chosen most frequently while answering questions in Subtest 1 (M=2.29, SD=2.59), followed by Subtest 2 (M=2, SD=1.38). Subtest 3 and Subtest 4 had lower mean values (M=1.41, SD=1.34, and M=1.67, SD=1.52, respectively).

In the third place was L1, *single sentence*, i.e. the participants found the answer within a single sentence. It was mostly in Subtest 3 that the participants marked this option in the protocol form (M=1.74, SD=1.49), followed by Subtest 4 (M=1.2, SD=.84), Subtest 1 (M=1.1, SD=1.27) and Subtest 2 (M=.98, SD=1.09).

L4, *without read* and L5-*no answer* had the lowest mean scores. The mean scores of L4 ranged between 0.02 - 0.19, and those of L5 ranged between 0.13 - 0.57.



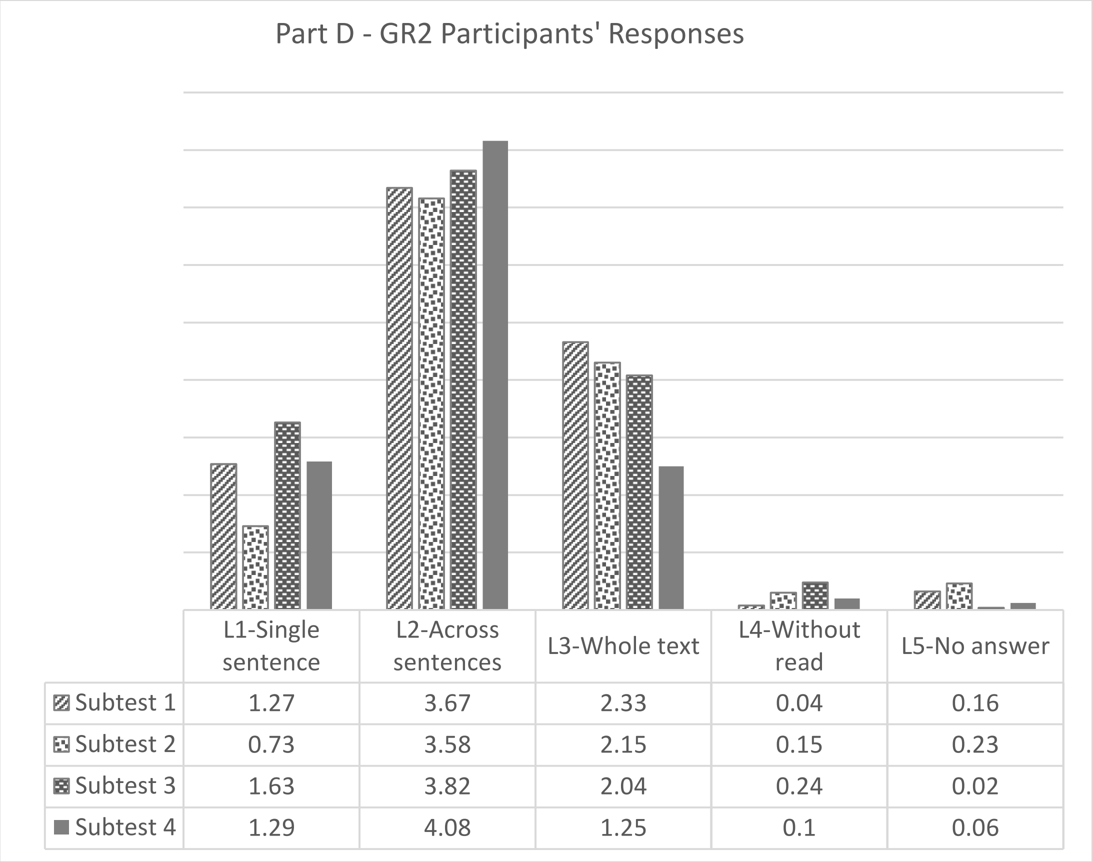
	L1-Single sentence	L2-Across sentences	L3-Whole text	L4-Without read	L5-No answer
▨ Subtest 1	1.1	3.69	2.29	0.08	0.25
⊡ Subtest 2	0.98	3.04	2	0.06	0.57
▩ Subtest 3	1.74	4.17	1.41	0.19	0.13
■ Subtest 4	1.2	3.56	1.67	0.02	0.49

Figure 24 GR1 participants' responses to Part D

4.3.1.1.5.2 GR2 participants' responses. As can be seen in Figure 25, among the five choices (L1-L5), L2, *across sentences* had the highest mean scores. In Subtest 4, L2 mean was the highest (M=4.08, SD=1.84), followed by Subtest 3 (M=3.83, SD=1.53), Subtest 1 (M=3.67, SD=2.74) and Subtest 2 (M=3.58, SD=1.58).

Next, L3, *whole text*, was the second most frequent choice by GR2 participants with mean scores ranging between 1.25 – 2.33: Subtest 1 had the highest mean (M=2.33, SD=2.39) and Subtest 4 had the lowest mean (M=1.25, SD=1.21).

The third most frequently chosen option was L1, *single sentence*. The mean scores were between 0.73 – 1.63, with Subtest 3 having the highest mean (M=1.63, SD=1.06), and Subtest 2 having the lowest (M=0.73, SD=0.68).



	L1-Single sentence	L2-Across sentences	L3-Whole text	L4-Without read	L5-No answer
Subtest 1	1.27	3.67	2.33	0.04	0.16
Subtest 2	0.73	3.58	2.15	0.15	0.23
Subtest 3	1.63	3.82	2.04	0.24	0.02
Subtest 4	1.29	4.08	1.25	0.1	0.06

Figure 25 GR2 participants' responses to Part D

L4, *without read* and L5, *no answer* had the lowest mean scores among the five options. L4 mean scores ranged between 0.04 – 0.24, and L5 mean scores ranged between 0.02 – 0.23.

Comparing the results from GR1 and GR2 participants' protocol forms, both groups stated that they used L2, *across sentences*, most frequently: mean scores from both groups were higher than 3.00. The lowest means scores were in L5, *no answer* (M<0.50), suggesting that the majority of the respondents thought they found the answer to the questions.

4.3.1.2 Summary of results. In this first version of the reading test, the participants filled out a retrospective protocol form by marking three different parts of it after answering each reading comprehension question. The first part asked about previewing strategies. The analysis of all responses revealed that each answer (*I read* *the text slowly and carefully*, *I read the text quickly and selectively,* and *I did not read the text*) were chosen with similar frequencies. However, an analysis of GR1 and GR2 participants' (according to the level groups they were placed at the DBE) responses revealed that the former preferred to read the text *quickly and selectively* more than GR2 participants did. In addition, GR2 participants showed a preference to skip the text (*I did not read the text*) and read the questions first.

The second part of the protocol form focused on test response strategies. The participants reported that they used four of the strategies more frequently than the others: S8, *read slowly & carefully*, S9, *re-read parts, S2, search & match similar*, and S6, *use vocabulary*. As reading carefully and re-reading parts of the text were the most frequently used strategies, it suggests that the participants tried to understand the text or parts of it to arrive at an answer. S2, *search and match similar* points to a strategy that was often preferred by GR1 participants. Rather than decoding the meaning in the text to arrive at an answer, GR1 participants matched words in the question with words in the text to find the location of the answers, and perhaps the answer to the questions.

The third part of the protocol form aimed to reveal whether the participants used local or global reading while responding to the questions. Overall, the participants reported that they read across sentences to find the answer, which points to global reading. There was local reading and whole text reading to some extent, but much less.

4.3.2 Introspective investigation. The second version of the reading test consisted of four different texts and questions related to each. There was a total of 30 items again, with one-point weight for each item. The compilation of the second version is given in Table 26.

Table 26 The compilation of the second version of the reading test

Version 1	**Version 2**
Subtest 1 (discarded)	Subtest 5 (new)
Subtest 2	Subtest 4 v2
Subtest 3 (discarded)	Subtest 2 v2
Subtest 4	Subtest 6 (new)

Two of the subtests (1 and 3) from Reading Test V1 were discarded due to insufficient item parameter values (for a discussion of this please see Section 3.3.3). Two new reading subtests replaced them: Subtests 5 and 6. The former contained a reading text on sea animals, and eight items related to it. Six of the questions tested macro level reading (i.e. understanding the main point of a paragraph) and two were multiple choice items testing critical reading (i.e. inferencing). In Subtest 6 there was a text about a seed preservation facility with eight open-ended items which required the participants to do *search reading* (i.e. a reading type which is a very basic type of reading to locate and understand discrete pieces of text) (Enright et al., 2000). Search reading is frequently used by undergraduate students (Urquhart & Weir, 1998). Two subtests (Subtest 2 and 4) from Reading T V1 were kept but some revisions were made in the items.

4.3.2.1 Data analysis. Despite the fact that the data collected was qualitative, it was noticed during coding that the transcriptions included strategies and processes that were repeatedly articulated by the participants. Quantifying the reported processes provided another angle in presenting a clear overall view of the strategies and skills that were used. Quantitative analysis involved calculating the frequency rates of the use of strategies.

4.3.2.1.1 Participant grouping. The participants in this second phase of the study consisted of students from the advanced, upper-intermediate and intermediate level groups, only. When analyzing the data, it was hypothesized that the participants' proficiency levels could have an impact on the choice of strategies used while answering the questions. For this reason, the data was first analyzed as a whole, then separately for participants belonging to one of the two groups: high-scoring (who received a score of 23 or more) or low-scoring (who received a score lower than 18). These groups were obtained by rank ordering the participants from the highest scoring to the lowest, and roughly dividing the group into three. As expected, the participants from the Advanced group (7) at the DBE made up the high-scoring group (except for two), and mainly the participants from the intermediate group and upper-intermediate group (8) made up the low-scoring group.

4.3.2.1.2. Strategy coding. Strategy coding was carried out following the coding rubric from Cohen and Upton's (2006) investigation on the strategies used in the new TOEFL reading test. Their approach in identifying a strategy is revealed in the following definition: "[It is] a specific and recognizable strategic choice made by the subject that is deliberate and purposeful and is intended to facilitate the reading or test-taking task" (p.39).

In the analysis of the verbal data, similar to Cohen and Upton (2006), the length of the strategy units were not pre-defined; rather, those strategies that were openly referred to were coded no matter how long they were. Usually, though, sentences, sometimes phrases or even words pointed to strategy use (e.g. "what's this word?", "It doesn't fit.").

The strategy rubric, adapted from the aforementioned study had the following structure:

1) Reading Strategies (RS)
   a) Approaches to reading the passage
   b) Uses of the passage and the main ideas to help in understanding
   c) Identification of important information and the discourse structure of the passage
2) Test Management Strategies (TM)
3) Test Wiseness Strategies (TW)

The analysis of these strategies provided information about reader's interaction patterns with the text, and the effect of their strategy choice on their comprehension of the text (Cohen & Upton, 2007).

During coding, some verbalizations, specific strategic choices, which did not fit with any of the codes in the Cohen and Upton rubric were added as new codes. Under the group *Approaches to reading the passage*, there are two new strategy codes, under *Test management strategies* there are eight new strategy codes and under Test Wiseness Strategies there is one new strategy code. The list of the rubrics is in APPENDIX G (new codes are added in italics).

4.3.2.2 A summary of strategy use across items and question types.

Table 27 presents a summary of the frequencies of strategy use for each item type, and for each of the strategy category.

The item types were categorized as follows:

Item focus	Item type	Expected reader purpose
Vocabulary (VOC)	Multiple Choice (MC)	Careful local
Macro level comprehension (MALC)	MC	Careful local and global
Search reading (SR)	Short Answer (SA)	Search reading Careful local and global
Expeditious reading (ER)	Matching (MAT)	Expeditious reading

The reading strategy that had the highest frequency rates (FR) across all items, except VOC was RS6 (MALC FR=16.50, SA FR=19.38, and MAT FR=12.33) (for VOC items, TM5 mean was higher: FR=11.00), which is a careful reading strategy. This result reveals that for most of the items, the participants carried out careful reading. In addition, the verbal reports revealed that the participants used scanning at a similar rate to careful reading (M=19.3) while they were answering SA questions. Another reading strategy that was used frequently across items was RSNEW2, *re-reading the text* (VOC FR=9, MALC FR=7.43, SR FR=4.38, and MAT FR=3.67).

The test management strategies that were used frequently across the three item types was TM5, *reads the question and then reads the passage/portion to look for clues to the answer, either before or while considering options*, (VOC FR=11.00, MALC FR=11.93 and SR FR=17.25). TM4, *reads the question and considers the options before going back to the passage/portion*, and TM22, *selects options through vocabulary, , sentence, paragraph, or passage overall meaning*, were used frequently across two item types (TM4: VOC FR=7.50, MALC FR=9.71, and TM22: VOC FR=6.50, MALC FR=6.57).

The strategies that were used exclusively in SR items were TMNEW2, *identifies answer through vocabulary, sentence, paragraph, or a number of paragraphs' overall meaning*, TMNEW3, *identifies section relevant to the question based on content*, TMNEW4, *identifies section relevant to the question: uses keywords*, TMNEW5, *identifies section relevant to the question: uses discourse structure*, and TMNEW6, *identifies section relevant to the question: uses subtitles*. Those strategies were specific to search reading purposes and were used in similar rates – except for TMNEW2, which had the highest rate among all (TMNEW2 FR=5.13, TMNEW3 FR=3.13, TMNEW4 FR=2.75, TMNEW5 FR=2.00, TMNEW6 FR=3.13).

Table 27 Frequency rates for all strategies across different item types

		Careful Local Reading - VOC (MC)	**Careful Global Reading - MALC (MC)**	**Search Reading (SA)**	**Expeditious Global Reading (MAT)**
	Reading strategies				
RS1	Plan a goal	0,50	0,57	0,13	2,00
RS2	Make a mental note	0,50	0,71	0,13	0,50
RS4	Read text carefully	-	0,14	-	0,50
RS6	Read a portion carefully	10,50	16,50	19,38	12,33
RS7	Scan	2,00	3,21	19,38	2,33
RS8	Look for markers of meaning	0,50	0,64	1,63	0,33
RS9	Repeat, paraphrase	2,50	3,21	0,88	3,33
RS10	Identify unknown word	2,00	2,14	0,25	2,67
RS11	Identify unknown sentence	0,50	0,07	0,25	0,33
RS12	Reread	1,50	0,86	2,13	1,00
RS13	Ask overall meaning	2,50	1,36	0,75	1,67
RS14	Monitor understanding	0,50	1,64	0,88	1,83
RS15	Adjust comprehension (previous)	-	1,29	0,88	1,67
RS16	Adjust comprehension (new)	0,50	2,64	1,63	3,83
RS17	Confirm understanding	1,50	2,57	1,38	1,67
RS19	Identify keyword	-	0,86	0,50	2,33
RS20	Search main idea	0,50	0,50	0,88	1,50
RS21	Use discourse knowledge	-	0,57	0,63	0,67
RS22	Use organization knowledge	-	0,71	0,75	0,33
RS23	Use logical connectors	-	0,14	0,38	0,17
RS24	Read ahead	-	1,00	0,75	0,33
RS25	Go back	3,50	1,50	0,90	0,17
RS26	Verify referent	0,50	0,36	0,13	-
RS27	Infer meaning (internal)	-	0,50	0,13	0,17
RS28	Infer meaning (external)	3,00	0,14	0,13	0,67
RSNEW1	Skim	1,50	1,14	2,75	2,00
RSNEW2	Read text again	9,00	7,43	4,38	3,67

Table 27 Continued.

		Careful Local Reading - VOC (MC)	Careful Global Reading - MALC (MC)	Search Reading (SA)	Expeditious Global Reading (MAT)
	Test management strategies				
TM1	Reread question	1,00	5,64	19,38	0,50
TM2	Paraphrase question	-	0,43	2,38	-
TM3	Wrestle with question intent	-	0,36	1,75	0,33
TM4	Read the question & options	7,50	9,71	0,13	2,50
TM5	Read the question & text	11,00	11,93	17,38	1,67
TM6	Predict own answer after reading	1,00	0,43	0,13	0,33
TM7	Predict own answer before reading	0,50	0,43	-	0,17
TM9	Identify unknown vocabulary	0,50	0,14	-	-
TM11	Consider a familiar option	-	0,29	-	0,83
TM12	Select option though uncertain	3,00	1,14	-	1,83
TM15	Drag the option to the sentence	1,50	1,07	-	-
TM17	Wrestle with option meaning	0,50	0,79	-	0,17
TM18	Make a guess	0,50	0,71	0,25	0,17
TM20	Locate vocabulary in context	5,50	0,21	-	-
TM22	Select option (meaning)	6,50	6,00	0,13	4,17
TM24	Select option (meaning/elimination)	3,50	6,64	-	0,90
TM25	Select option (elimination)	-	0,50	-	-
TM26	Select option (discourse)	-	1,00	0,13	-
TMNEW1	Identify answer (keyword)	-	-	0,88	-
TMNEW2	Identify answer (meaning)	-	-	5,13	-
TMNEW3	Identify section (content)	-	-	3,13	-
TMNEW4	Identify section (keyword)	-	-	2,75	-
TMNEW5	Identify section (discourse)	-	-	2,00	-
TMNEW6	Identify section (subtitles)	-	-	3,13	-
TMNEW7	Identify keywords in Q	0,50	0,93	8,25	1,17
TMNEW8	Identify unknown vocabulary	1,50	0,71	1,50	0,33
	Test wiseness strategies				
TW1	Elimination	0,50	0,57	-	-
TW3	Select by keyword	-	0,29	-	1,00
TWNEW1	Use item sequence information	-	-	0,75	-

4.3.2.3 Detailed analysis of strategy use. In this section, for each of the question types (careful local reading – vocabulary, careful global reading – macro level comprehension, search reading, and expeditious global reading) tables with usage frequencies and frequency rates are presented. The frequencies were calculated by dividing the number of strategy use into the number of questions of that type, and the strategies were rank ordered from the highest value to the lowest. Those strategies with a mean lower than 1 were not included in the table. In order to demonstrate how the participants vocalized strategy use, examples are given. At the end of each exemplary transcription the participant's level group and the line numbers for the utterance in the transcript file are also given. As all the participants were Turkish, I translated all the utterances from Turkish to English. The lines that were found to be important in revealing the use of the specific strategy are underlined.

4.3.2.3.1 Detailed analysis of strategies for vocabulary items (Multiple Choice). The vocabulary items intended to measure the test takers' ability to guess the meaning of a word or a phrase using contextual clues. These items were prepared as careful local reading items, as the answer to the questions could be found within a sentence in the reading text.

Table 28 presents the most commonly used strategies for this item type, the frequency of use and the use ratios. The expected strategies were careful local reading of the sentence that contains the unknown vocabulary, and if necessary reading the preceding and following sentence to be able to guess its meaning from context. The chosen vocabulary were low frequency items, that is, they did not occur in the language commonly. The purpose was to have the students understand the context of the sentence to be able to guess what that specific word means. It was anticipated that the vocabulary items could not be answered through world knowledge or background knowledge.

Table 28 Strategies used in answering vocabulary items

Code	Strategy Name	Frequency	Rate
Reading Strategies			
RS6	Reads a portion of the text carefully.	21	10.5
RSNEW2	Reads the whole text/paragraph one more time carefully	18	9
RS25	Uses other parts of the text to help in understanding a given portion: Goes back in the text to review/understand information that may be important to the remaining text.	7	3.5
RS28	Infers the meanings of new words by using work attack skills: External context (neighboring words/sentences/overall passage).	6	3
RS9	Repeats, paraphrases, or translates words, phrases, or sentences—or summarizes paragraphs/whole text—to aid or improve understanding.	5	2.5
RS13	During reading asks self about the overall meaning of the whole text/portion.	5	2.5
RS10	Identifies an unknown word or phrase.	4	2
Test Management Strategies			
TM5	Reads the question and then reads the passage/portion to look for clues to the answer, either before or while considering options.	22	11
TM4	Reads the question and considers the options before going back to the passage/portion.	15	7.5
TM22	Selects options through vocabulary, sentence, paragraph, or passage overall meaning (depending on item type).	13	6.5
TM20	Looks at the vocabulary item and locates the item in context.	11	5.5
TM24	Selects options through elimination of other option(s) as unreasonable based on paragraph/overall passage meaning.	7	3.5
TM12	Considers the options and selects preliminary option(s) (lack of certainty indicated).	6	3

Note: Low frequency items (LF) (<2.00) are excluded from this table.

4.3.2.3.1.1 Reading strategies. The reading strategy that was used most frequently was RS6 (FR=10.5). Despite the fact that these items were careful local reading items, the participants commonly used careful global reading, i.e. they read two or more sentences, and tried to connect information from them to arrive at an understanding. Examples of this strategy are:

1. [*Reads paragraph*.]
   Actually, [the question] is not about the meaning of the paragraph. It's asking 'it pays'. It would suffice if I read the previous sentence and the following sentence.
   [*Reads paragraph one more time*.]
   Oh, it means it is worth it. Let's read the options.
   [*Reads options*.]
   Here it means, it brings benefits. That is, it is beneficial to do that.
   (P-AD3: 253 – 258)

2. [*Reads item 18*.]
   We're going to read paragraph D. I'll read this single sentence. I don't have to read the whole paragraph. There I found 'it pays'.
   [*Reads sentence*.]
   What does the preceding sentence say? I hope it is not too long.
   [*Reads sentence*.]
   [*Reads options*.]
   I guess, it requires curiosity.
   I am going to read the last sentence again.
   [*Rereads sentence*.]
   I see, when it says pays, it means it costs…that is, it costs or it helps to learn the warning signs.
   [*Rereads options*.]
   It says 'pays', I know 'pay attention', but 'it pays' is not something I use a lot. 'It pays', it is like 'it requires'. That's how I feel. I think the answer is 'b', it requires curiosity.
   (P-AD8: 210 – 223)

3. [*Reads paragraph*.]
   'It pays to learn the warning signs'.
   I didn't understand, I'll read again.
   [*Rereads paragraph*.]
   Is he talking about himself, or is he trying to tell the reader something? No, no. 'It pays to learn the warning signs.' Gosh. Is it about the whole sentence? I need to read the other paragraphs, perhaps I'll get it then.
   (P-AD1: 93 - 95)

4. [*Reads item 10*.]
   It is asking the word 'humbug'.
   I think the answer is 'b' but I guess I will read the paragraph. It's been underlined. I'll start from the previous sentence.
   [*Reads text.*]

It is not like what I thought. It might be ‘c’. I’ll do this later.
I’ll read the whole text.
[*Reads text from the start again and then summarizes what s/he reads*.]
Here, I thought ‘humbug’ was a lie. I couldn’t understand exactly what it means here.
(P-AD5: 37 - 47)

The second most frequently used reading strategy was one of the new codes added to the strategy rubric: RSNEW2 (FR=9). The need for a new code for this strategy occurred because most of the time the participants explicitly explained that they were going to read / reading the text one more time. Here is an example:

5. [*Reads paragraph*.]
   I’ll read the last four sentences of this paragraph again.
   [*Rereads a portion of paragraph*.]
   Now I’m looking at the options.
   [*Rereads paragraph. Translates as s/he reads*.]
   It is not ‘c’. I am in between two choices. I didn’t exactly understand what ‘b’ says but I feel it is the answer. It is not ‘promotional tactic’ because the text says they did something they shouldn’t have done.
   (P-UI4: 83 – 90)

The next most frequently used reading strategy, though only used for about one-third of the time in comparison to RS6, was a RS25 (FR=3.5). The participants went back in the text to understand information that may be important to understand a section or the question. Here are two examples:

6. [*Reads the question, then reads sentence with the words “it pays”*.]
   Let’s take a look at the previous sentence.
   [*Reads sentence*.]
   What can replace this?
   [*Reads options*.]
   [*Rereads sentence*.]
   (P-AD2: 122 – 128)

7. Let’s read the paragraph quickly.
   [*Reads paragraph carefully*.]
   It has nothing to do with “A”. It brings benefits. Now, in the previous sentence it says …
   [*Reads sentence*.]

That's why the answer is "c".
(P-UI5: 216 - 219)

Another reading strategy used in guessing word meaning was RS28 (FR=3). The participants used information from the external content, i.e. neighboring words, or sentences, to infer the meaning of a new word. Here is an example:

8. [*Reads the question*.]
I don't know the word 'humbug'. First I'm going to read the sentence with 'humbug'.
[*Reads text carefully*.]
It should be cheating.
[*Reads options*.]
But I don't know the meaning of 'fraud' so I'm going to look at the text again.
[*Reads text quickly*.]
Probably, it has nothing to do with 'a'. I eliminated 'c'. It should be 'b'.
(P-UI5: 115 – 122)

The next reading strategy used while answering vocabulary items was RS9 (FR=2.5). The participant either repeated, paraphrased or translated words or sentences to improve her understanding of the text. Here is an example of this strategy use:

9. Let's see, what is this about 'humbug'. I'm reading the options.
[*Reads options.]*
I don't know the meaning of 'fraud'. I guess it is something like stealing. It should be something negative.
[*Rereads text*.]
Hmmm. Let's translate this into Turkish.
[*Rereads, and translates*.]
[*Rereads options*.]
Oh, I got it.
(P-AD8: 117 – 118)

To a lesser extent, the participants used RS13: they asked themselves about the overall meaning of the whole text/portion.

10. [*Reads the question.]*
[*Reads the paragraph carefully.]*
I am somewhat confused here because the way they describe the introverts here … is odd.
(P-AD5: 93 – 95)

A number of participants used RS10 (FR=2), which refers identifying an unknown word or phrase.

11. [*Reads the question*.]
[*Reads the paragraph and translates into Turkish.]*
So they could access information that looked like deleted.
[*Reads options.*]
I don't know the word 'fraud', I couldn't understand 'b'. Can it b 'a'? No.
[*Skims text.*]
It cannot be a virus. It does not talk about that kind of virus. What is 'humbug'? I'll read the other sentence.
[*Reads sentence.]*
I get a meaning like cheating. I am trying to understand the meaning of 'promotional tactic'. It doesn't fit with the meaning of it Turkish 'promosyon'. It cannot be a tactic to promote something. 'Humbug' looks like a type of behavior. Even though I don't know the word 'fraud' I choose this option.
(P-IN5: 112 – 121)

4.3.2.3.1.2 Test Management Strategies. The most popular strategy in answering vocabulary items was TM5 (FR=11). In answering the questions, the majority of the participants read the question, and then read the text to find the answer either before or while considering options. Here is an example:

12. [*Reads the question.]*
[*Reads the paragraph carefully*.]
Let's look at the options.
[*Reads the options*.]
[*Rereads the sentence with the word in question*.]
'It pays' means it will bring you benefits. I mark option 'c'.
(P-AD6: 173 – 178)

The next most frequent test management strategy was TM4 (FR=7.5). This time, the participants read the options before going back to the text.

13. [*Reads the question*.]
It is asking the meaning of 'humbug'.
[*Reads the options*.]
I think the answer is 'b', but I'm going to read the paragraph. It has already been underlined. I'll start reading from the previous sentence.
(P-AD5: 37 – 38)

The third most frequently used test management strategy was TM22 (FR=6.5). This strategy refers to choosing an option through vocabulary, sentence, or passage overall meaning. Here are two examples of the use of this strategy:

> *14.* [*Reads the question.*]
> I'm going to look at the options first.
> [*Reads the options.*]
> [*Reads the sentence that contains the word 'humbug'.*]
> They didn't delete the profiles, so it is an act of fraud.
> (P-AD2: 60 – 64)

> *15.* [*Reads the question.*]
> [*Reads paragraph D carefully.*]
> [*Rereads the question and the options.*]
> I don't think it is 'c' because the paragraph doesn't say anything about it's benefits. Let's replace this word with these options.
> [*Replaces 'it pays' with option 'a'.*]
> Yes, it is 'b' because any one of the people you see everyday may be an introvert, and you make him/her angry. You need to be curious to understand that that person is an introvert.
> (P-UI3: 104 – 111)

The next test management strategy used while answering vocabulary items was TM20 (FR=5.5). The participants looked at the vocabulary item and located it in context. Here is an example of the use of this strategy:

> *16.* [*Reads the question.*]
> I need to find this 'it pays'. OK, the last sentence. I need to read the whole paragraph.
> [*Reads the paragraph.*]
> I think it means, it helps.
> (P-IN9: 156 – 159)

Sometimes the participants chose an option through the elimination of other options, which is TM24 (FR=3.5) in the list of test management strategies. Here is an example of the use of TM24:

> *17.* [*Reads the question and the options.*]
> I don't think it is 'a', because the text talks about a disrespect towards those people but… now it says 'attackers' … I will read the options again.
> [*Reads the options.*]
> I don't think it is a computer virus. The best option is 'b'.
> (P-IN9: 69 – 71)

Another test management strategy is TM12 (FR=3). After reading the question, some participants chose a preliminary option without being too certain. Here is an example:

> *18.* [*Reads the question.*]
> [*Reads the options*.]
> I'm going to read the last four sentences of this paragraph again.
> [*Reads sentences*.]
> I'm looking at the options again.
> [*Reads the options*.]
> [*Rereads the paragraph*.]
> [*Translates part of the text.*]
> It is not 'c'. I'm in between 'a' and 'b'. I don't understand what is meant in 'b' but it feels like this one is the answer. It is not *promotional tactic* because it says they did something they shouldn't have done. I am not sure.
> (P-UI4: 83 – 90)

The examples given here reveal that the strategies were not used in isolation: some strategies were consistently used together. For example, the participants commonly read (RS6) and reread (RSNEW2) the related sentences, or even the whole paragraph if they failed to understand the meaning conveyed. Here is an example of those two strategies used together:

> 19. [*Reads sentence*.]
> What does this mean?
> [*Rereads sentence*.]
> 'Require curiosity', 'attract attention', these are too close. But they do not fit in the sentence. 'Attract attention the learn the warning signs."
> No, it is not good. I say, 'bring benefits'.
> (P-IN5: 220 - 224)

T24, the elimination strategy to arrive at an answer, was generally used together with RS6, reads a portion of the text carefully. Here is an example:

> 20. Oh, this....I think this is the word 'humbug', I need to find out what it means. What are the options?
> [*Reads options*.]
> I'll read the sentence again.
> [*Rereads sentence containing 'humbug'.]*
> What does it say?
> [*Rereads paragraph*.]
> What was announced as deleted was actually broadcasted.
> [*Rereads options*.]
> It doesn't make sense.

> I'll mark 'an act of fraud' for the time being.
> (P-AD3: 127 – 137)

4.3.2.3.2 Detailed Analysis of Strategies for Macro Level Comprehension Items (Multiple Choice). The items that require macro level comprehension were marked as careful global reading items according to the new reading model. In this question type, the test takers were expected to separate main ideas from supporting details, understand how an argument develops throughout the text, distinguish generalizations and examples, and make inferences (see Table 13). Critical reading questions were also dealt with under careful global reading. In responding to these items, the test takers were expected to be deeply engaged in the text, and analyze the text, or interpret information not given explicitly in the text. As such, the reader was expected to read carefully and derive meaning by understanding the relation between sentences; therefore, it is usually global reading. The frequencies and rates for the macro level comprehension items are given in Table 29.

4.3.2.3.2.1 Reading strategies. As the macro level comprehension and critical reading items necessitate, *reading a portion of the text carefully* (RS6) has the highest frequency rate (FR=16.50), followed by reading the question. The frequency rate of RS6, in this section is much higher than that in the vocabulary section (16.50 and 10.50, respectively). Here are two examples:

1. This is the last paragraph. I remember there was a question related to this. I need to understand what it means so I am going to read slowly and try to understand it all.
   (P-AD1: 15 – 15)

2. I'm moving on to paragraph E. I quickly looked at the question; it is about meaning, that's why I'm going to read paying attention to the meaning of the paragraph.
   (P-AD5: 98 – 98)

Similar to the order of the strategy use, RSNEW2 was the second most frequently used strategy (FR=7.43). After reading a portion of the text, the participants reread to improve their understanding of the text, or to monitor that they have understood it right. Here are two examples:

3. [*Reads paragraph.*]
   At the end of this sentence, he talks about the same type of people.
   [*Reads options*.]
   I couldn’t understand, I’m going to read this last part again.
   [*Rereads part of paragraph*.]
   (P-IN9: 136 – 140)

4. [*Reads the question.*]
   [*Reads paragraph.*]
   I couldn’t understand this part. I’m going to read again.
   [*Reads paragraph and translates into Turkish.*]
   (P-IN3: 166 – 171)

Table 29 Strategies used in answering macro level comprehension items

Code	Strategy Name	Frequency	Rate
Reading Strategies			
RS6	Reads a portion of the text carefully.	231	16.50
RSNEW2	Reads the whole text/paragraph one more time carefully	104	7.43
RS7	Reads a portion of the text rapidly looking for specific information.	45	3.21
RS9	Repeats, paraphrases, or translates words, phrases, or sentences—or summarizes paragraphs/whole text—to aid or improve understanding.	45	3.21
RS16	Adjusts comprehension of the text as more is read: Identifies the specific new information that does or does not support previous understanding.	40	2.64
RS17	Confirms final understanding of the text based on the content and/or the discourse structure.	26	2.57
RS10	Identifies an unknown word or phrase.	30	2.14
Test Management Strategies			
TM5	Reads the question and then reads the passage/portion to look for clues to the answer, either before or while considering options.	167	11.93
TM4	Reads the question and considers the options before going back to the passage/portion.	136	9.71
TM24	Selects options through elimination of other option(s) as unreasonable based on paragraph/overall passage meaning.	93	6.64
TM22	Selects options through vocabulary, sentence, paragraph, or passage overall meaning (depending on item type).	92	6.57
TM1	Goes back to the question for clarification: Rereads the question.	79	5.64

Note: Low frequency items (LF) (<2.00) are excluded from this table.

Scanning, RS7 [FR=3.21], was also used despite the questions being careful reading items. RS7 was used as a strategy to eliminate or chose an option. The participants scanned for words in an option in the text, and read that part of the text to help them eliminate or chose the option. Here are some examples:

5. [*Reads the question.*]
   [*Reads options and translates each option.*]
   [Reads the paragraph.]
   [*Returns to the options.*]
   It is not 'a' because it says they were dominating. There's 'abnormal' in 'b'. I don't know what that means, so if I cannot find it in the text, then there's a problem.
   [*Scans text for abnormal.*]
   I couldn't find it in the text. But, normal, *anormal*, abnormal… They sound similar.
   (P-IN3): 286 – 295)

6. [*Reads the question*.]
   Now I'm going to search for 'university students'.
   [*Scans text and locates 'university students'*.]
   I'll read this sentence from the start.
   (P-IN6: 94 – 99)

7. [*Reads the question.*]
   'University students' is my keyword.
   [*Scans text for 'university students'.*]
   Right there, at the beginning of the paragraph.
   [*Reads paragraph carefully*.]
   P-IN3: 190 – 195

With the same frequency as RS7, RS9 was used (FR=3.21). The participants used this strategy to help their understanding of the text by repeating, paraphrasing, or translating words or phrases. Here are two examples:

8. The last part may contain a clue about the writer's attitude, but I couldn't get it.
   [*Rereads part of the paragraph*.]
   [*Translates a sentence*.]
   There is an ironical expression here; this might be critical. If there is irony, it may be approving as well. But, I think it is critical.
   (P-IN3: 200 – 205)

9. [*Reads the question*.]
   Let's look at the questions first.
   [*Reads paragraph*.]
   [Translates paragraph.]
   [*Continues reading*.]

I am going back to the options now.
[*Reads the question and options*.]
[*Rereads the paragraph*.]
[*Translates paragraph*.]
[*Rereads the options*.]
(P-IN9: 133 – 143)

RS16 is related to monitoring the consistency of the text as more is read. The participant checks whether the new information she receives is consistent with the meaning representation of the text she has constructed so far. The frequency rate of RS16 was 3.21. Here are two examples:

10. As far as I can understand, developed countries wanted to be a part of this because the seeds come from their own country. As a result of the discussions it talks about a postponement.
    (P-AD4: 270 - 270)

11. Oh, I see, they only attack those they want to eat, they don't attack for the sake of attacking. At least, I understood this much.
    (P-AD8:41 - 41)

Another reading strategy used in a similar proportion to RS16 was RS17 (FR=2.57). Using the content or the discourse structure, the participants confirm their final understanding of the text. Here are some examples:

12. I think, deletion of personal accounts was a lie. Because the hackers revealed them…the deleted information. It means they did not keep their words.
    (P-IN3: 109 – 109)

13. [*Reads paragraph.*]
    I see, the real scandal was this. The answer is 'c'. They are not honest about they delete accounts after users want. And hackers showed it.
    (Talked in English.)
    (P-UI6: 45 – 47)

RS10 was used in MALC items with a frequency rate of 2.14. It refers to the participants' identification of an unknown word or phrase while reading the questions or the text. Here are two examples:

14. [*Reads the question.*]
    I don't know 'refute. I'm checking again. I'll skim a bit.

[*Skims paragraph*.]
I didn't really understand. He does not support. Because I don't know what 'refute' means, To be on the safe side I'll say he is questioning.
(P-AD7: 82 – 85)
[*Skims text.*]
I don't know what 'indifferent to' means. But I guess it is something like neutral.
(P-AD2: 104 – 105)

4.3.2.3.2.2 Test management strategies. In answering MALC items, TM5 (FR=11.93) was the most frequently used test management strategy followed by TM4 (FR=9.71). The former strategy refers to reading the question and then reading the text, the latter refers to reading the question and then reading the options before going back to the text. Here are examples for these two strategies:

15. [*Reads question.]*
It says here, their behavior is limited. Why is that so?
[*Rereads text.*]
Oh, ok.
[*Reads options.*]
Now, I should check which one is logical.
[*Rereads option A.*]
(P-AD8: 80 – 86)

16. [*Reads question.]*
I'm going to read again because I didn't understand.
[*Rereads question*.]
What is the writer's attitude?
[*Reads options*.]
It's in the last paragraph.
[*Reads paragraph*.]
(P-IN5: 173 – 179)

17. [*Reads question and options*.]
[*Reads paragraph A carefully*.]
Here, it talks about an introvert. They like to spend time by themselves, so the answer is 'b'.
(P-UI1: 111 – 113)

In the third and fourth places are TM24 and TM22, with frequency rates very close to each other. TM24 (FR=6.64) is about selecting options through elimination of other options as unreasonable based on paragraph meaning and TM22 (FR=6.57) is about selecting options through vocabulary, sentence, or paragraph overall meaning. Here are some examples of TM24 use:

18. It does not say anything about thinking before talking. So I cross ‘b’.
    Here too, it talks about something negative. But in option ‘c’ it says they are considerate. That’s why it is not possible. I choose ‘a’.
    (PUI-1: 140 - 140)
19. [*Reads option A.]*
    There is nothing about it here.
    [*Reads option B.*]
    This might be.
    [*Reads option C.]*
    It doesn’t say anything about this.
    (P-AD2: 31 - 36)

20. I eliminated 3 because there is no ‘new technology’. It has nothing to do with potential threat. So I choose ‘7’.
    (P-IN5: 63 - 63)

Here are examples of TM22 uses:

21. [*Reads question 14 and the options*.]
    Now I’m going to read the paragraph from the start and then when I get that point I’ll be super careful.
    [*Reads paragraph*.]
    I didn’t understand; I’ll read that sentence again.
    [*Reads sentence*.]
    Now I’ll try to choose from the options.
    (P-IN9: 88 – 93)

22. [*Reads options*.]
    There’s nothing here about this. They do research; they consider this important. I think it is ‘a’ because it says the UAV’s could not find what they were looking for.
    (P-IN7: 47 - 47)

23. Now I’ll try to choose from the options. I think it is the first one: the paragraph talks about things they should avoid, that there are some problems. Now, option ‘a’ covers all this.
    (P-IN9:93 - 93)

The last test management strategy is TM1 (FR=5.64), that is going back to the question for clarification. Here is an example:

24. I’m going to read the question again.
    [*Rereads question*.]
    Which option provides this?
    [*Rereads options*.]
    (P-AD4: 179 – 182)

4.3.2.3.3 Detailed analysis of strategies for matching heading items. Although, these items were grouped under the title matching, and the aimed reading strategy was primarily skimming, i.e. expeditious reading, the most frequently used strategy was careful reading. The items in this category required the test takers to understand the gist or the main idea of each paragraph and match them with one of the headings given in the options. Apparently, to understand the main idea, in some of the paragraphs, the participants needed to read carefully than expeditiously to match each option with a heading.

As can be seen in Table 30, the most frequently used reading strategy was RS6 (FR=12.33) followed by RS16 (FR=3.83) though it was used only about one-third of the time compared to RS6.

In the following example, the participant skims the text first (expeditious reading); failing to find the answer, reads the paragraph carefully.

1. [*Skims paragraph*.]
   [*Reads paragraph carefully*.]
   I don’t know the meaning of this, but since the writer uses ‘but’ here…
   I’ll try to understand it … I’ll read the previous sentence.
   (P-IN8: 38 – 41)

In this example, the participant notices that the answer was in the first line of the paragraph. If he had read the first and last lines of the paragraph, it’s highly probable that he could have found the answer:

2. <u>I’ll read paragraph C starting from the middle of the paragraph because there is some unnecessary information at the beginning.</u>
   [*Reads paragraph C carefully*.]
   ‘Prey’ is a keyword.
   [*Continues reading*.]
   I didn’t understand much but this is not about researching animals.
   What is this about?

[*Reads the question again.*]
It is about food.
[*Continues reading.*]
It talks about feeding habits. I made a mistake in skipping the first part of the paragraph.
(P-UI2: 46 – 55)

Table 30 Strategies used in answering matching items

Code	Strategy Name	Frequency	Rate
Reading Strategies			
RS6	Reads a portion of the text carefully.	74	12.33
RS16	Adjusts comprehension of the text as more is read: Identifies the specific new information that does or does not support previous understanding.	23	3.83
RSNEW2	Reads the whole text/paragraph one more time carefully	22	3.67
RS9	Repeats, paraphrases, or translates words, phrases, or sentences—or summarizes paragraphs/whole text—to aid or improve understanding.	20	3.33
RS10	Identifies an unknown word or phrase.	16	2.67
RS19	Identifies and learns the keywords of the text.	14	2.33
RS7	Reads a portion of the text rapidly looking for specific information	14	2.33
RSNEW1	Skims text	12	2.00
RS1	Plans a goal for the text.	12	2.00
Test Management Strategies			
TM22	Selects options through vocabulary, sentence, paragraph, or passage overall meaning (depending on item type).	92	4.17
TM4	Reads the question and considers the options before going back to the passage/portion.	136	2.50

Note: Low frequency items (LF) (<2.00) are excluded from this table.

The participants used RSNEW-2 (FR=3.67), reading the text/paragraph one more time, carefully very frequently when answering matching questions. This strategy refers to returning to the text to reread after a failed attempt at answering a question, or lack of certainty about the chosen answer.

3. First, I'm reading the questions because the paragraph is long, I need to choose the questions. There are matching questions. Ok. We're going to pick out what is in each paragraph.
   [*Reads option* 1.]
   White sharks. Let's see where we can find 'white sharks.'
   [*Scans paragraph A*.]
   Not this one. I'll keep going.
   [*Skims paragraph B*.]
   It doesn't say anything about what it feeds on.
   [*Scans paragraph C*.]
   Ok. We understand what the white sharks feed on here.
   (P-UI6: 2 – 12)
   I underlined the words; now, I'm searching for them in the text.
   (P-AD5: 23 – 23).

The reading strategy RS9 (FR=3.33) was used to improve or aid understanding of the text. Here is an example of RS9 use in matching items:

4. [*Reads paragraph E.]*
   [*Reads options*.]
   It is about how they were interested when they saw the AUV.
   [*Reads paragraph F.]*
   [*Reads options*.]
   This part is about how they were frustrated because they did not find what they were looking for.
   (P-AD6: 35 – 40)

A common approach towards responding the matching questions was to read the paragraphs first and then to eliminate the options. Some participants used keywords to make the matching; the strategy RS19 (FR=2), identifying and learning the key words of the text, was used with this type of item. Here are examples of the reporting of the use of RS19:

5. 'Attack' may be a keyword here.
   (P-IN3: 36 – 36)

6. I underline 'motivation'.
   (P-AD2: 17 – 17)

7. It says ‘unlike’ in the last sentence. That’s a keyword.
   (P-AD6: 46 – 46)

8. I found the name Ashley Madison. I’m underlining it.
   (P-UI4: 63 – 63)

Using the strategy RS19 (FR=2.00), *Identifies and learns the key words of the text*, the participants noticed some important words in the text and used them as clues to reach some information. Here are some examples:

9. In the second paragraph it talks about how we observe them. Now, I’ll read the third paragraph.
   [*Reads paragraph*.]
   I underlined the part where it explains how they feed.
   (P-AD3: 33 – 34)

10. It talks about white shark. Then, the topic is white shark
    (P-IN5:4 – 4)

11. I’m quickly reading the text to see if there are any keywords.
    (P-AD6: 12 – 12)

Some participants used RS7 (FR=2.33), *reads a portion of the text rapidly looking for specific information*, with the intention of finding words that are similar to the keywords in the question. Some examples of this strategy are:

12. I’m reading paragraph A. Blah, blah, blah.
    [*Scans text*.]
    It’s not here. I’ll read paragraph C.
    [*Scans text*.]
    Blah, blah, blah. Next lines. It says…so and so. Actually, it talks about their observations.
    (P-AD8: 61 – 67)

13. ‘Recorded media.’ Where did I see it? I’m looking.
    [*Scans text*.]
    (P-AD8: 47 – 47)

RSNEW1 was used with a frequency rate of 2.00. The participants skimmed the text to get an overall idea on what each paragraph was about.

14. First the text…I'm going to read the first paragraph, then the first and last sentences of the others, so I'll know what kind of information there is in each paragraph. Then, I'll read the questions.
    (P-IN5: 2 – 3)

15. First I'm going to read the questions.
    [Quickly reads questions.]
    I think these are matching questions. Now, I am trying to figure out the keywords. I'm looking at the paragraphs to understand what each is about. I'm going to read the first sentences of each.
    (P-IN6: 2 – 3)

Another strategy used with matching items was RS1 (FR=2.00), *plans a goal for the text*. The participants' verbalizations on how they decide to proceed with the text were coded with this strategy. Here are some examples:

16. First, I check the question types.
    [*Quickly reads the questions*.]
    Now, there is matching, then multiple choice questions. In order to match them I need to read the whole text.
    (P-IN4: 5 – 5)

17. First, I'm going to the text…I'm going to read the first paragraph. Then, I'll read the first and last sentences of the other paragraphs so that I'll have an idea as to what information can be found where.
    (P-IN5: 2 – 2)

18. I look at the questions. I see that there is one for each paragraph. I will first read the paragraph and then try to get the overall meaning.
    (P-UI2: 3 – 3)

4.3.2.3.3.2 Test management strategies. Among the test management strategies, two of them had a frequency rate over 2: TM22 (FR=4.17) and TM4 (FR=4.50). TM22 is about selecting options through meaning (of vocabulary, sentence, or paragraph). The participants mentioned how the meaning they obtained from a portion of a text matched with the question intent, or how a vocabulary item used in the text coincided with some vocabulary in the text. Here are some examples:

19. Option 'c' is reasonable. It does not talk about a myth like password security and then, they ask for money. I say 'c'.
    (P-IN5: 111 – 111)

20. I'm going to choose 'c' because the writer says introverts should not be treated this way, and there are also criticisms, so I choose 'c'.
(P-UI1: 151 - 151)

21. I think 'b' because I matched 'misconceptions' and misunderstood'.
(P-UI2: 168 - 168)

22. What I understand from the whole of the paragraph is that the impact team did not do this to mean harm to people. That's why I say 'c'.
(PUI3: 56 - 56)

TM4 refers to reading the question and considering the options before going back to the text. In answering the matching items, the participants used this strategy either to identify the keywords of the paragraphs, or to get an idea from the options as to what to expect from each paragraph beforehand. Here are two examples:

23. [*Reads the question.*]
First I'm going to read these options so that I have an idea beforehand about the paragraphs.
[*Reads options.*]
Alright, so the text is about sea animals, now I'm going to read the text.
(P-AD5: 3 – 6)

24. First, I'm reading the questions.
[*Reads the question and options.*]
This is about paragraph headings, so I'll go to the text and read each paragraph.
(P-IN7: 3 – 4)

4.3.2.3.4 Detailed analysis of strategies for search reading items. Search reading refers to searching for a pre-determined topic in a long text and reading carefully to understand the relevant part. The search reading part is presented in Reading Test V2 contains a text longer than that in the other parts, and eight short answer questions. The participants were expected to search the text for the topic of each question and then read carefully to find the answer. (Table 31).

4.3.2.3.4.1 Reading strategies. The most frequently used strategy in this section was RS6 (FR=19.75), careful reading, followed by RS7 (FR=19.38). After reading the question, the participants scanned the text to locate a keyword they identified in the question, or locate a section which they believed was relevant to the question. In case they couldn't find the exact keyword from the question, they went back to the question to identify other keywords or the topic - hence the use of a test-management strategy, TM1. As mentioned earlier, search reading questions were developed with the aim of having the test taker read carefully across sentences to be able to arrive at an answer. The data showed that the most frequently used strategy was RS6, confirming the item writers' expectations.

The existing strategies from the Cohen and Upton rubric were found to be insufficient in explaining test taker behavior especially in the search reading part of the test. During the analysis of the data, eight new test management strategies were defined, appropriate with the item type, six of which were used frequent enough to be included in Table 31. The first two of the new strategies, TMNEW1 and TMNEW2 are related to the approach in finding the answer, either through a keyword or meaning derived from the words, sentences, or paragraphs. TMNEW3 to TMNEW6 are related to the identification of the section of the text where the participants thought the answer is located. TMNEW7 is identifying keywords in the question and TMNEW8 is identifying unknown vocabulary in the question.

Table 31 Strategies used in answering search reading items

Code	Strategy Name	Frequency	Rate
Reading strategies			
RS6	Reads a portion of the text carefully.	158	19.75
RS7	Reads a portion of the text rapidly looking for specific information.	155	19.38
RSNEW2	Reads the whole text/paragraph one more time carefully	35	4.38
RSNEW1	Skims text	21	2.75
RS12	During reading rereads to clarify the idea.	17	2.13
Test management strategies			
TM1	Goes back to the question for clarification: Rereads the question.	153	19.13
TM5	Reads the question and then reads the passage/portion to look for clues to the answer, either before or while considering options.	138	17.25
TMNEW7	Identifies keywords in the question	66	8.25
TMNEW2	Identifies answer through vocabulary, sentence, paragraph, or passage overall meaning	41	5.13
TMNEW3	Identifies section relevant to the question based on content	25	3.13
TMNEW6	Identifies section relevant to the question: uses subtitles	25	3.13
TMNEW4	Identifies section relevant to the question: uses keywords	21	2.75
TM2	Goes back to the question for clarification: Paraphrases (or confirms) the question or task.	19	2.38
TMNEW5	Identifies section relevant to the question: uses discourse str	16	2.00

Note: Low frequency items (LF) (<2.00) are excluded from this table.

The strategies RS6, and RS7 and RSNEW1 were frequently used together as the participants read the question, then read the text both carefully and expeditiously, and if they felt they could not locate the answer, reread the question to identify new keywords, or to establish the topic again. Here are examples of these strategies:

> 25. [*Reads question 30 and scans text for a keyword.*]
> If I look at the second paragraph, there is no mention of a threat. Let's look at the third paragraph.

[*Skims text.*]
It talks about gene banks here, but there is nothing about the risks that threaten them. I continue reading but there’s nothing related to a risk.
[*Scans text.]*
Here, I think it talks about it, the biggest threat, and those risks are lack of resources and funding.
[*Writes answer.*]
(P-UI1: 219 – 227)

26. [*Reads question 25.*]
    I continue reading the paragraph. I try to find the words administration of seed, decision, conflict.
    [*Scans text.*]
    It can be in the section Who owns the world’s heritage.
    [*Skims text.]*
    It can be here, in this part because it talks about treaty. Yes, it talks about ministry of agriculture. I think I’m close to the answer.
    [*Reads text.*]
    [*Rereads question 25.*]
    Recommended. Here it is. The answer is a chamber should be built inside a mountain.
    (P-IN7: 197 – 201)

27. [*Reads question 24.]*
    Chambers is like a room. The answer can be around here because it says description.
    [*Scans text.]*
    I found chambers.
    [*Reads text.*]
    It says this serves, so it must be talking about its function.
    [*Writes answer.*]
    (P-IN3:318 – 322)

The participants used RS12 (FR=2.13) to improve their understanding of the text. Here is an example:

28. [*Reads text and rereads question.]*
    [*Rereads text.*]
    Rules need to change. This does not fit with the question. I’ll read again.
    [*Rereads text.*]
    It is here, but I couldn’t spot it.
    [*Rereads question.*]
    Ok, something needs to be done.
    They need advance approval.
    [*Rereads text.*]
    I’ll skip this. Next question.
    (P-AD6: 310 – 321)

4.3.2.3.4.2 Test management strategies. The participants frequently reread the question for clarification. This strategy, TM1 (FR=19.38), was used in all item types but most frequently with search reading items. The reason for that is they were expected to identify the location of the answer by understanding the topic relevant to the question. Most participants, though, reread the question and scanned the text for related words, rather than locating the area of the text by matching the question topic with subtitles in the text. Here are examples of the use of this strategy along with TM5 (FR=17.25) which was usually used together with TM1, and TMNEW7 (FR=8.25) which is about identifying the keywords in the question:

29. [*Reads question 24*].
[*Rereads question 24.*]
The keywords are 'chambers' or 'specific layout'. So I'm looking at the section 'Description of the facility'. I'm scanning to find 'chambers'.
[*Scans text.*]
Here I found 'layout of the chambers'. What was the question?
[*Rereads question 24.*]
[*Reads text.*]
[*Rereads question 24.*]
'Layout is purposeful', but why is it specific? It's asking the reason.
[*Rereads question 24.*]
I should look at the previous paragraph, perhaps the answer is there.
Or I should read the part after 'purposeful'.
(P-UI2: 228 – 236)

30. [*Reads question 24*].
I'll try to find 'chambers'. I'm scanning now.
[*Scans text.*]
Here, I found 'chambers'. I'll read the question again.
[*Rereads question 24.*]
[*Reads text.*]
Is the answer 'purposeful'?
[*Skims text.*]
[*Reads text carefully.*]
[*Rereads question 24.*]
Because…
[*Reads text.*]
'This serves as a security measure.' Is this the answer?
(P-UI4: 224 – 235)

Three of the test management strategies TMNEW3 (FR=3.13), TMNEW4 (FR=2.75) and TMNEW6 (FR=3.13) were used to identify the section of the text relevant to the question. Use of TMNEW3 shows identification of location through content, use of TMNEW4 shows identification of location through keywords and use of TMNEW6 shows identification of location through subtitles. Here are examples to the use of these three strategies:

31. [*Reads question 24*.]
    I need to find something related to 'chambers'.
    [*Reads subtitle, 'Description of the facility'.]*
    It should be here.
    [*Reads text*.]
    (P-UI1: 160 – 163)

32. [*Reads question 24*.]
    This is not related to the introduction section. It talks about the purpose, why it was established. This must be in the description section.
    (P-AD6: 239 – 240)

33. [*Reads question 27*.]
    I'll find this in 'Why Svalbard?' section.
    [*Rereads question 27.]*
    Ok, 'Why Svalbard?' is the right place.
    [*Reads text.]*
    (P-UI6: 202 – 207)

Here are examples of the new strategy TMNEW2 (FR=5.13) which was used frequently throughout the search reading section to identify the answer through vocabulary, sentence, paragraph or a number of paragraphs' overall meaning. Here are two examples:

34. [*Reads text*.]
    I want to read the first paragraph, from the start very carefully.
    [*Reads paragraph carefully*.]
    Here it mentions the seed bank but there is no mention of a purpose.
    There is nothing in the previous sentences that I can pick clearly.
    [*Continues reading*.]
    I guess this one, preserving from seeds. At first I thought that it was talking about what they were doing and it did not mention the purpose but now this preserving seems like a purpose.
    (P-AD5: 124 – 129)

35. [*Reads question 25*.]
    'Administration of seed' may be after this paragraph, so I'll keep reading.
    [*Continues reading*.]
    I don't think the answer is here, so I'll skip this part. I'll search for 'administration of seed'. Then I'll read around that part. Now I'll scan.
    [*Scans text*.]
    [*Rereads question 25*.]
    I missed the question.
    [*Rereads question 25*.]
    I'm going to scan whole page till I find it.
    [*Scans text*.]
    It says something here. I'll read the question again because I found 'Administration of seed'. Now, I'm searching for the 'decision'. It says 'heated debate'. So they discussed. I'm looking for the decision.
    [*Scans text*.]
    It says 'postponed'. It was postponed, but that isn't a decision. I might read the other paragraph.
    [*Continues reading*.]
    Not here. This part is on something different. I should find it in the first paragraph.
    (P-AD7: 195 – 214)

4.3.2.4 Summary of the results of detailed analysis of strategies. Overall, RS6 (FR=14.68) was the most frequently used strategy across all item types. In comparison among the four item types, it was again the most frequently used strategy except vocabulary items. In answering vocabulary questions, the test takers initially read only the sentence that contained the lexical item in question. They might have already known the meaning of the lexical item that was being asked, in which case, they could mark an option without reading any further. Or, they might have guessed the meaning of the word from the sentence where it was located – although the test writers did their best to avoid this. Most of the time they had to use other strategies such as reading the preceding and following sentences. This is a plausible reason why the use of RS6 in responding vocabulary questions was not as frequent as in responding other item types.

All of the participants approached the reading text as a test; therefore, predominantly they read the question and the options (TM4, FR=4.96) before viewing or reading the text. A more frequent strategy, though, was after reading the text before reading the options (TM5, FR=10,96).

In dealing with the vocabulary questions, the participants had a similar approach: first, they read the question and then they read part of the text carefully Whereas the majority of the participants tried to find the answer by reading the text first TM5, FR=11), some others looked at the options first (TM4, FR=7.5). Many participants marked an option through vocabulary, sentence, or paragraph meaning (TM22, FR=6.5). Nonetheless, there were others who chose an option through elimination of other options as unreasonable (TM24, FR=3.5).

While answering matching items, rather than reading expeditiously to obtain a general idea about the gist of the paragraph, the participants preferred to read carefully (RS6, FR=12.33). Reading carefully refers to reading with the intention of understanding the main ideas and supporting details, understanding both explicit statements and implied meanings, understanding the macro structure of the text, understanding relations between ideas and arguments, etc. The decision to read carefully rather than expeditiously might have been due to text organization: perhaps the paragraphs were not structured enough to provide a clear picture of the gist. In expeditious reading, common reading approach is reading the first and last sentences of paragraphs, randomly reading words to map the text, or to read the first and last paragraphs of the text.

The second most frequently used test management strategy was used in only about one fourth of the time compared to reading carefully: *selecting options through meaning* (TM22, FR=4.17). Apparently, while matching items, some participants preferred to use the clues in the options to arrive at an answer (TM4, FR=2.50), rather than discovering the gist of the passage themselves (TM5, FR=1.67). Some of them matched the keywords in the options with those in the paragraphs to help them choose an option (TMNEW7, FR=1.17). Other commonly used strategies were mainly reading strategies, such as RS16 (FR=3.83), that suggested the use of some higher level cognitive processes: integrating what is read recently to what has been read before (Field, 2004). The participants used another reading strategy, RS9 (FR=3.33), which suggested the use of monitoring and remediation, e.g. the participants translated sentences into their L1, and sometimes paraphrased what they've read to aid their own understanding.

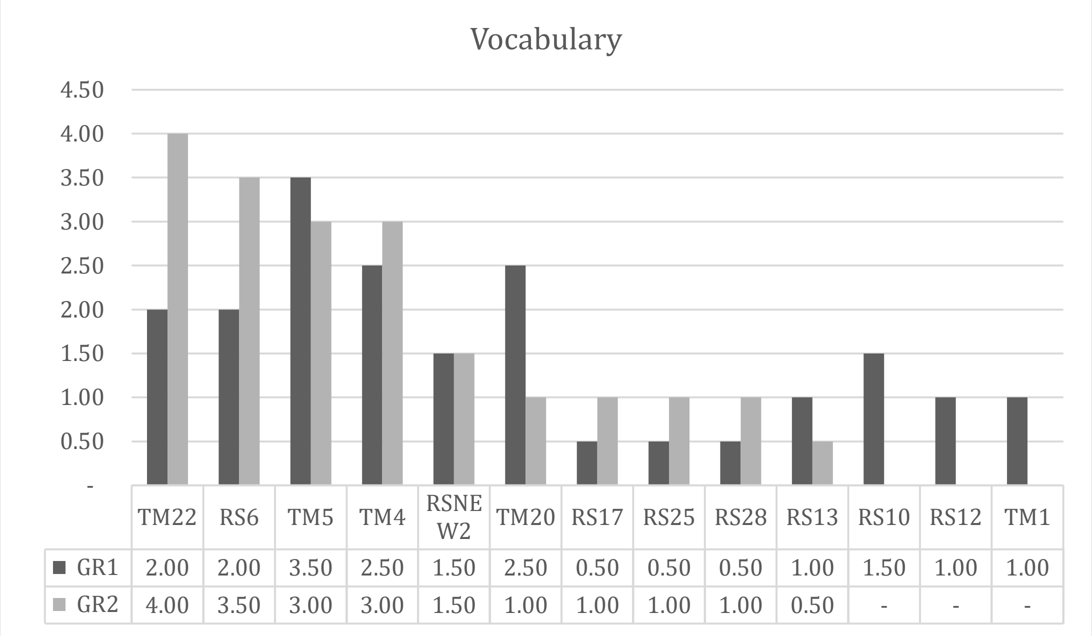
Macro level comprehension and critical reading questions required the participants to read portions of text carefully. This was the most frequently used strategy (RS6, FR=16.50) followed by reading part of a text to find clues to the answer before or while considering the options (TM5, FR=11.93). Again, approaching the text with the intention to answer the questions, many participants first read the question and looked at the options, and then read the text (TM4, FR=9.71). Three test management strategies were popular with this type of questions: choosing an option through elimination of other options (TM24, FR=6.64), choosing an option through meaning (TM22, FR=6.57), and rereading the question for clarification (TM1, FR=5.64). To a lesser extent, the participants also used scanning (RS7, FR=3.21).

Due to the nature of some of the questions, such as inference or attitude questions, the participants used a number of higher level processes such as confirming that the new information they obtain from the text is congruent with what they have read previously (RS16, FR=2.64), and paying attention to the discourse structure of the text (RS17, FR=2.57).

The search reading questions were different from the rest as they were in the short answer format. Naturally, the processes that were activated during the participants' attempts at answering those questions somewhat differed from those of the selected response items. The participants used three strategies in same frequencies, reading carefully (RS6, FR=19.38), scanning (RS7, FR=19.38), and reading the question a second time (TM1, FR=19.38). The short answer format led the respondents to identify keywords in the question (TMNEW7, FR=8.25), and (TMNEW2, FR=5.13) scan for those keywords in the text. When they found keyword, they started reading carefully. Due to the fact that the texts used in search reading were much longer than the others, identifying the section relevant to the question in a short time (TMNEW3, FR=3.13, TMNEW4, FR=2.75, TMNEW5, FR=2.00, and TMNEW6, FR=3.13) was a necessary strategy to be able to complete this part of the test in time.

4.3.2.5 Comparison of high- and low-scoring participants' use of strategies.

As it was assumed that the participants' level of proficiency in English could have an effect on their choice of strategies, the participants were sorted from high to low according to the score they received from the reading test, and the choice of strategies from the top and the bottom one third of participants were compared.



	TM22	RS6	TM5	TM4	RSNEW2	TM20	RS17	RS25	RS28	RS13	RS10	RS12	TM1
■ GR1	2.00	2.00	3.50	2.50	1.50	2.50	0.50	0.50	0.50	1.00	1.50	1.00	1.00
■ GR2	4.00	3.50	3.00	3.00	1.50	1.00	1.00	1.00	1.00	0.50	-	-	-

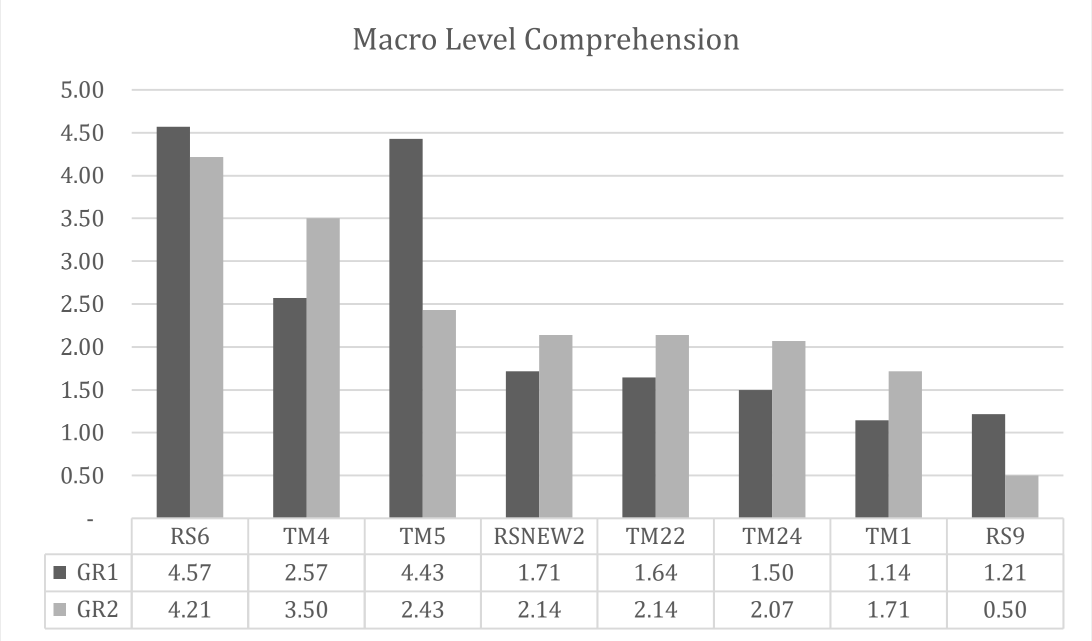
TM22	Selects options through vocabulary, sentence, paragraph, or passage overall meaning.
RS06	Reads a portion of the text carefully.
TM5	Reads the question and then read the passage/portion to look for clues.
TM4	Reads the question and considers the options before going back to the passage /portion.
RSNEW2	Reads the whole text/paragraph one more time carefully
TM20	Looks at the vocabulary item and locates the item in context
RS17	Confirms final understanding of the passage based on the content or discourse structure
RS25	Uses other parts of the text to help in understanding a given portion: Goes back in the text to review/understand information that may be important to the remaining text.
RS28	Infers the meanings of new words by using word attack skills: External context (neighboring words, etc.).
RS13	During reading asks self about the overall meaning of the whole text/portion
RS10	Identifies an unknown word or phrase
RS12	During reading rereads to clarify the idea
TM1	Goes back to the question for clarification: rereads the question

Note: The strategies that had frequencies lower than 1 in both groups were excluded from the graph.

Figure 26 Strategy use in answering VOC items by GR1 and GR2

4.3.2.5.1 Vocabulary items. According to Figure 26, in answering the vocabulary questions, compared to the low-scoring participants the strategies that were used about twice as more frequently by the high-scoring participants were TM22 (FR=4), RS6 (FR=3.5), RS17 (FR=1), RS25 (FR=1), and RS28 (FR=1). High-scoring participants' strategies were geared more towards making meaning of what they were reading as the use of the strategies reveal. Low-scoring participants, on the other hand, used the strategies TM5 (FR=3.5), TM20 (FR=2.5), RS13 (FR=1), RS10 (1.5), RS12 (FR=1), and

TM1 (FR=1) more frequently than the high-scoring participants did. The low-scoring participants' use of test management strategies more frequently than high-scoring participants may suggest a compensation mechanism for their lack of proficiency.



	RS6	TM4	TM5	RSNEW2	TM22	TM24	TM1	RS9
■ GR1	4.57	2.57	4.43	1.71	1.64	1.50	1.14	1.21
■ GR2	4.21	3.50	2.43	2.14	2.14	2.07	1.71	0.50

RS06	Reads a portion of the text carefully.	TM22	Selects options through vocabulary, sentence, paragraph, or passage overall meaning.
TM4	Reads the question and considers the options before going back to the passage/portion	TM24	Selects options through elimination of other options as unreasonable based on paragraph/overall passage meaning
TM5	Reads the question and then reads the passage to look for clues to the answer, either before or while considering options	TM1	Goes back to the question for clarification: rereads the question
RSNEW2	Reads the whole text/paragraph one more time carefully	RS9	Repeats, paraphrases or translates words, phrases or sentences to aid or improve understanding

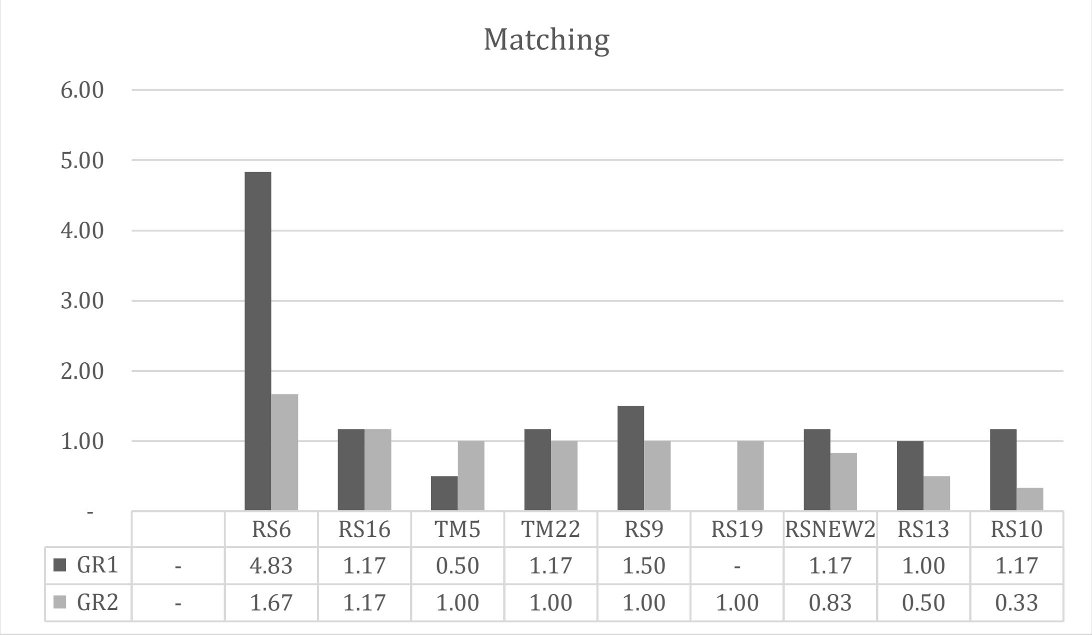
Note: The strategies that had frequencies lower than 1 in both groups were excluded from the graph.

Figure 27 Strategy use in answering MALC items by GR1 and GR2

4.3.2.5.2 Macro level comprehension and critical reading items. The comparison of high-scoring and low-scoring participants' scores in macro level comprehension items revealed that the most frequently used strategy for both groups was RS6, *reading a portion of the text carefully*, that is, careful global reading (low-scoring participants' FR=4.57, high-scoring participants' FR=4.21) (Figure 27). As these items intended to assess test takers' ability in understanding main ideas, separating main ideas from supporting details, and the links between macro propositions in the text, the use of this strategy seems fitting. The strategy, TM4, *reading the questions and then reading the passage or a portion of the text to look for clues*, was the second most frequently used strategy for the high-scoring group (FR=3.50) followed by TM5 (FR=2.43), *reading the question and then reading the passage to look for clues to the answer, either before or while considering options*. It is not surprising that the most frequently used strategies by the high-scoring group were all related to careful global reading. In contrast, for the low-scoring participants, the most frequently used strategy was RS6 (FR=4.57), followed by TM5 (FR=2.57), a strategy again related to careful global reading.

An interesting difference between the high and low-scoring participants' use in reading strategies was in TM4, *reading the question and then reading the passage or a portion of the passage to look for clues.* This strategy was used at a much lower frequency by the low-scoring participants (high-scoring participants' FR=3.50, low-scoring participants' FR=2.57). Rather than going back to the passage to look for clues, this group of participants preferred to read the options beforehand.

Another strategy that were used differently by the two groups of participants was RS9, *repeating paraphrasing or translating words, phrases or sentences to aid or improve understanding*. The low-scoring participants use the strategy about twice as much as the high-scoring participants did (FR=1.21).



		RS6	RS16	TM5	TM22	RS9	RS19	RSNEW2	RS13	RS10
■ GR1	-	4.83	1.17	0.50	1.17	1.50	-	1.17	1.00	1.17
■ GR2	-	1.67	1.17	1.00	1.00	1.00	1.00	0.83	0.50	0.33

RS6	Reads a portion of the text carefully.
RS16	Adjusts comprehension of the text as more is read: Identifies the specific new information that does or does not support previous understanding.
TM5	Reads the question and then reads the passage to look for clues to the answer, either before or while considering options
TM22	Selects options through vocabulary, sentence, paragraph, or passage overall meaning.
RS9	Repeats, paraphrases or translates words, phrases or sentences to aid or improve understanding
RS19	Identifies and learns the keywords of the text
RSNEW2	Reads the whole text/paragraph one more time carefully
RS13	During reading asks self about the overall meaning of the text
RS10	Identifies an unknown word or phrase

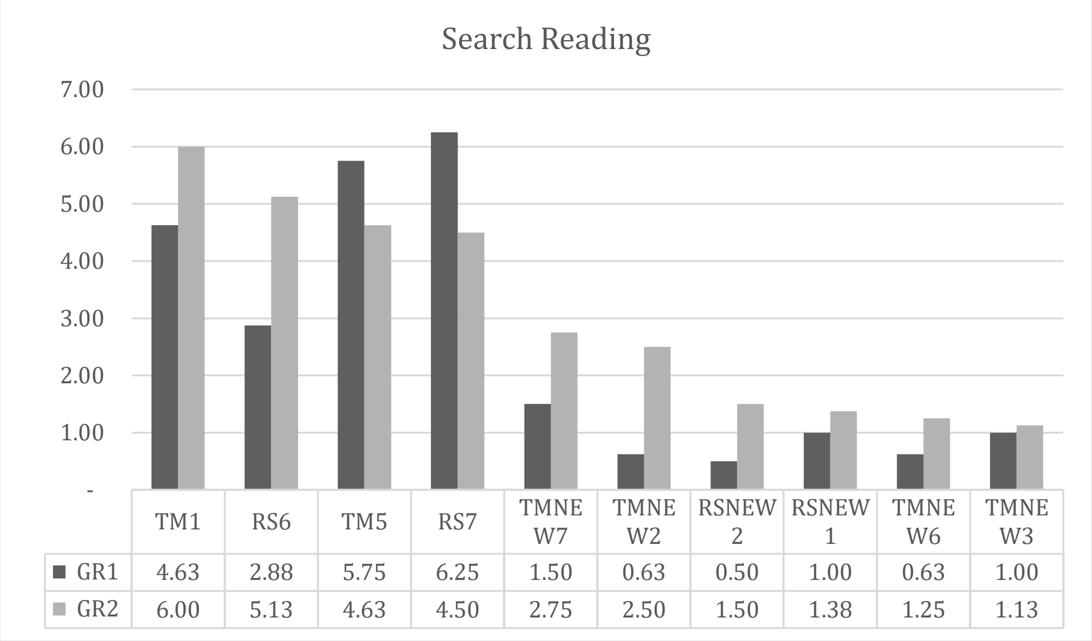
Note: The strategies that had frequencies lower than 1 in both groups were excluded from the graph.

Figure 28 Strategy use in answering MAT items by GR1 and GR2

4.3.2.5.3 Matching items. Matching items asked the participants to match each paragraph with a heading from a list of options (Figure 28). The most frequently used strategy when responding these items was RS6, *reading a portion of the text carefully*. However, the participants in the low-scoring group used this strategy about three times more than the others (low-scoring participants' FR=4.83). All the other strategies related to the matching items were used at low frequencies by both participant groups except for, RS 19, *identifying and learning keywords of the text*, which was used only by the high-scoring participants (FR=1.0).

4.3.2.5.4 Search reading items. Figure 29 shows that in search reading, the most frequently used strategies by GR2 were TM1 (FR=6.00), RS6 (FR=5.12), and TM5

(4.62). The most frequently used strategies for GR1 were different: this group used RS7 (FR=6.25), followed by TM5 (FR=5.75). The biggest difference between the two participant groups was in the use of TMNEW2: GR2 used the strategy about five times more than GR1 did (GR2 participants' FR=2.50, GR1 participants' FR=0.63).



	TM1	RS6	TM5	RS7	TMNEW7	TMNEW2	RSNEW2	RSNEW1	TMNEW6	TMNEW3
■ GR1	4.63	2.88	5.75	6.25	1.50	0.63	0.50	1.00	0.63	1.00
■ GR2	6.00	5.13	4.63	4.50	2.75	2.50	1.50	1.38	1.25	1.13

TM1	Goes back to the question for clarification: rereads the question	TMNEW2	Identifies answer through vocabulary, sentence, paragraph, or a number of paragraphs' overall meaning
RS06	Reads a portion of the text carefully.		
TM5	Reads the question and then reads the passage to look for clues to the answer, either before or while considering options	RSNEW2	Reads the whole text/paragraph one more time carefully
		RSNEW1	Skims text
RS07	Reads a portion of the text rapidly looking for specific information.	TMNEW6	Identifies section of relevant to the question: uses subtitles
TMNEW7	Identifies keywords in the question	TMNEW3	Identifies section relevant to the question based on content

Note: The strategies that had frequencies lower than 1 in both groups were excluded from the graph.

Figure 29 Strategy use in answering SR items by GR1 and GR2

4.3.2.6 Summary of the comparison between GR1 and GR2 participants. In general, the GR2 participants' use of reading and test management strategies were more appropriate for the specific item type. When answering matching items, GR2 participants used a number of strategies in similar ratios, between 0.33 – 1.67, (RS6, RS16, TM5, TM22, RS9, RS19, RSNEW2, RS13, and RS10) whereas GR1 participants primarily used careful reading (RS6, FR=4.83), and used the other strategies in smaller ratios (between 1.17 – 0). As matching items mainly necessitate some expeditious reading strategies, it seems, GR1 participants did not manage to select the relevant strategies to answer the questions of this type.

In macro level comprehension and critical reading items, the focus was on understanding the whole of the text in detail; hence, both groups used the reading strategy, reading carefully extensively. However, GR1 spent shorter time spans while reading the text and reverted to the question quite often, perhaps to remind themselves what to look for while reading carefully.

Vocabulary items necessitated the participants to understand the meaning of a word or a word cluster by using contextual clues. Therefore, they were expected to read a few sentences, for both questions, and understand the connection between those sentences to be able to arrive at the correct answer. This was mainly what GR2 participants did: they selected an option based on meaning (FR=4.00). GR1, on the other hand, read the question and used elimination of the options to arrive at an answer (FR=2.5) more frequently then choosing an option based on meaning (FR=1.00).

In search reading, after understanding the question, GR2 managed to identify the section relevant to the question using titles, subtitles, or content, and then they read carefully to find the exact answer. GR1 participants, on the other hand, relied on scanning, i.e. matching words in the question with words in the text, to find the relevant section of the text. However, not all questions included exact words from the text, in some there were paraphrases, and therefore, scanning would not have helped them identify the location of the answer. They probably resorted to careful reading to locate the answer.

4.4 RQ3: To What Extent Do Item Parameters Contribute to the Validity Claims of the Test?

This research question was designed to reveal whether the scores from the reading test provide meaningful statistical measures for claims of scoring validity. To this end a number of statistical analysis within the domain of CTT were carried out for both versions (V1 and V2) of the test.

4.4.1 Reading Test V1 results. The first version of the reading test was administered to participants not as a whole but in parts; therefore, for each subtest (i.e., Text I + 8 items, Text II + 7 items, Text III + 8 items, Text IV + 7 items) I conducted the analyses separately. The subtests and the number of participants who took the tests are given in Table 32.

Table 32 Participant numbers on subtests

V1	Subtest 1	Subtest 2	Subtest 3	Subtest 4
Items	8	7	8	7
Participants	101	97	105	97

4.4.1.1 Descriptive statistics. Frequency of scores for all subtests is given in Table 33. Accordingly, in Subtest 1 and Subtest 2, scores ranged between 0 and 7, in Subtest 3, scores ranged between 1 and 8, and in Subtest 4 scores ranged between 1 and 7.

Table 33 Frequency of scores

	Subtest 1		**Subtest 2**		**Subtest 3**		**Subtest 4**	
Score	Freq.	%	Freq.	%	Freq.	%	Freq.	%
0	2	2.0	3	3.1	1	1.0	-	-
1	16	15.8	6	6.2	1	1.0	6	6.2
2	24	23.8	21	21.6	2	1.9	15	15.5
3	21	20.8	26	26.8	4	3.8	14	14.4
4	23	22.8	22	22.7	6	5.7	27	27.8
5	5	5.0	14	14.4	20	19.0	16	16.5
6	7	6.9	4	4.1	27	25.7	15	15.5
7	3	3.0	1	1.0	31	29.5	4	4.1
8	-	-	-	-	13	12.4	-	-
Total	101	100.0	97	100.0	105	100.0	97	100.0

As can be seen in Table 33, the most frequent scores were in the range of 2-4 out of 8 for Subtest 1 and out of 7 for Subtest 2. More than 65% of the participants' scores were within that range for those subtests. Subtest 3 was negatively skewed, with the most frequent scores being in the range of 6 and 7, out of 8. In Subtest 4, the most frequent score was 4 out of 7 and the distribution of the scores was symmetrical around the score of 4.

Other descriptive statistics related to the scores of the subtests are given in Table 34. Accordingly, Subtest 3 was the easiest test with an average score of 5.9/8 (74%) and Subtest 1 was the most difficult with an average score of 3.0/8 (38%). Factors affecting the high mean score of Subtest 3 could be the six **Yes/No/Not Given** items which, as an item type, overemphasize the testing of isolated factual details, i.e. testing of explicit information that requires mainly lower level reading skills (as opposed to higher level skills such as inferencing). Another factor could be related to the topic of the reading text; it contained mainly factual information and the Flesch reading ease parameter (see Section 2.2.3.3.1.1 for details on Flesch) revealed that this was the easiest text among the four. Consequently, score distribution of Subtest 3 was negatively skewed (-1.19).

Table 34 Descriptive statistics for the four subtests

	Subtest 1	**Subtest 2**	**Subtest 3**	**Subtest 4**
Mean	3.0 (38%)	3.3 (47%)	5.9 (74%)	4.0 (57%)
Std. Error of Mean	0.16	0.15	0.15	0.16
Std. Deviation	1.61	1.43	1.57	1.58
Variance	2.6	2.0	2.5	2.50
Skewness	0.50	0.03	-1.19	-0.06
Kurtosis	-0.19	-0.16	1.97	-0.76
Alpha Coefficient	0.43	0.25	0.51	0.38
Range	7	7	8	6
Minimum	0	0	0	1
Maximum	8	7	8	7

The participants who took the first version of the test were from six different level groups. In order to reveal the distribution of scores among participants at different levels of language proficiency, the results are presented in two categories. Those students who were from the pre-intermediate level, pilot pre- intermediate level, and the repeat group were categorized as Group 1 (GR1) and those from the advanced, upper-intermediate and intermediate levels were categorized as Group 2 (GR2). The distribution of the scores within these two groups for each subtest is given in Table 35.

As can be seen in Table 35, for Subtest 1 and Subtest 2, the majority of the scores of GR2 clustered around the scores of 3 – 5, whereas for GR1, this range was 2 – 4. In Subtest 3, the majority of GR2 scores were in the range of 6 – 7, for GR1 this range was 5 – 7. In Subtest 4, the majority of GR2 scores were within the range of 4 – 6, whereas for GR1 group the range was around 2 – 4. Overall, GR2 scored higher than GR1 in all four subtests.

Table 35 Distribution of test scores (V1) among GR1 and GR2 participants

Score	**Subtest 1**		**Subtest 2**		**Subtest 3**		**Subtest 4**	
	GR1	**GR2**	**GR1**	**GR2**	**GR1**	**GR2**	**GR1**	**GR2**
0	-	-	3	0	1	0	-	-
1	2	0	5	1	1	0	6	0
2	15	1	15	6	2	0	12	3
3	11	13	15	11	4	0	9	5
4	14	7	9	13	6	0	14	13
5	8	15	2	12	11	9	4	12
6	2	3	0	4	13	14	1	14
7	0	7	0	1	11	20	0	4
8	0	3			5	8		

In Table 36, the mean scores for all the subtests for these two groups of participants and the percentage values of the mean scores are given.

Table 36 Means and percentages according to GR1 and GR2

	GR1 mean	Percentage	GR2 mean	Percentage
Subtest 1	2.33	29%	3.80	48%
Subtest 2	2.57	37%	3.94	56%
Subtest 3	5.39	67%	6.53	82%
Subtest 4	3.02	43%	4.80	69%

As expected, GR2 participants had more correct answers in all subtests than GR1 participants. The easiest test for both groups of participants was Subtest 3, with a mean score of 5.39 for GR2 and 6.53 for GR1. The hardest test, for both groups, was Subtest 1: GR2 participants received a mean score of 3.80 and GR1 participants scored 2.33 on average. The largest difference between the mean scores of the two groups of participants was in Subtest 4 and the smallest difference was in Subtest 3.

4.4.1.2 Item analyses: Item facility and item discrimination. The analyses of the items in accordance with CTT conventions are given in this section. Table 36 presents item facility indices (IF), item discrimination indices (*d*) and point biserial correlations (r_{pbi}) for the items in all four subtests.

4.4.1.2.1 IF indices. As a brief reminder, item facility values closer to 0.50 provide the widest scope of variation among the test takers (Alderson et al., 1995), therefore item developers prefer IF values to be as close to the mid-point as possible to discriminate better between higher and lower ability test takers. As has been discussed previously in detail in Section 3.3.3, the IF value for the reading test has been set to the range of 0.40 and 0.80; that is, it was planned to include items that 40 – 80% of the test takers can answer correctly.

4.4.1.2.2 Discrimination indices. The discrimination power of the items (*d*) are expected to be equal to or higher than 0.20 (Crocker & Algina, 1986), and the point biserial correlations (r_{pbi}) are expected to be higher than 0.30 (L. Cohen, Manion, & Morrison, 2007).

Subtest 1 had two types of items: matching and MC. As Table 37 shows, in Subtest 1, all the matching items were outside the expected IF range of 0.40 - 0.80 while the MC items had acceptable facility indices. Five of the matching items were more difficult and one item was easier than anticipated. In terms of discrimination indices, all eight items discriminated well between participants with higher and lower ability levels when the lower limit of the discrimination index is taken as 0.20. For point biserial correlations, all items had an acceptable level of correlation of correct items with the total scores.

The matching items that had IF values lower than the 0.40 limit were intended as expeditious reading items. The participants were expected to read each paragraph expeditiously to get the gist, and then match the headings given in the question with a paragraph. However, the retrospective protocols revealed that the majority of the participants read those paragraphs carefully rather than expeditiously to understand the main idea in each. This might have created a time problem for the participants, as reading all six paragraphs carefully probably required more time than they were given. Another explanation for the low item facility values could be related to the wording of the options in the questions. The statements in the options could have been better expressed for clarity of meaning.

As a result, the computations revealed a mean score of 43% (lower than the expected value of 60%). After discussions with the testing committee, rather than revising those six items in Subtest 1, it was decided that a new reading task be prepared in the second version of the reading test.

Table 37 Item analysis results

Item #	Item type	Item Facility (IF)	Discrimination Index (*d*)	Point-biserial correlations (r_{pbi})
Subtest I				
1	Matching	0.82	0.41	0.36
2	Matching	0.37	0.70	0.61
3	Matching	0.28	0.44	0.53
4	Matching	0.14	0.26	0.38
5	Matching	0.27	0.41	0.53
6	Matching	0.32	0.56	0.50
7	MC	0.44	0.56	0.34
8	MC	0.43	0.44	0.33
Subtest II				
1	MC	0.33	0.56	0.39
2	MC	0.47	0.52	0.37
3	MC	0.33	0.26	0.36
4	MC	0.75	0.30	0.42
5	MC	0.36	0.33	0.35
6	MC	0.43	0.15	0.28
7	MC	0.57	0.07	0.29
Subtest III				
1	Y/N/NG	0.89	0.17	0.27
2	Y/N/NG	0.76	0.62	0.55
3	Y/N/NG	0.85	0.45	0.42
4	Y/N/NG	0.87	0.38	0.43
5	Y/N/NG	0.59	0.31	0.34
6	Y/N/NG	0.92	0.21	0.27
7	MC	0.56	0.90	0.48
8	MC	0.85	0.24	0.31
Subtest IV				
1	MC	0.37	0.56	0.18
2	MC	0.51	0.70	0.37
3	MC	0.66	0.44	0.16
4	MC	0.65	0.52	0.32
5	MC	0.58	0.48	0.32
6	MC	0.61	0.67	0.31
7	MC	0.59	0.48	0.26

Notes: MC: Multiple Choice, Y/N/NG: Yes/No/Not Given. The cells outside the expected ranges are shaded in grey.

In Subtest 2, there were seven multiple choice items (MC). While four of the items had acceptable facility values (0.40 and above), three (i.e., items 1, 3, and 5) had facility values lower than 0.40. In addition, the discrimination indices of two of the items (items 6 and 7) were lower than 0.20. Those same items also had low point biserial coefficient values (see Table 35). On the other hand, the reading text in Subtest 2 afforded the inclusion of critical reading type items, such as inferencing and drawing conclusions, and the testing committee preferred to keep the text and the items with good parameters, and revise those items that had low item facility values and point biserial correlations.

In Subtest 3, there were eight questions. Six of them were Y/N/NG items and two were MC items. The results revealed that Y/N/NG items were easier than expected. Five of them had a facility index higher than 0.80 and the mean score of the test was 5.9 (74%). One item's discrimination index was off limits (< 0.20) and two of the items had point-biserial correlations lower than expected (<0.30). The committee decided to discard this subtest as well, since the Y/N/NG item types did not yield much information about the test takers as the items were easier than planned, and the text was mostly concrete (a text about sea animals), making it difficult to write more challenging items.

Subtest 4 had seven MC items and it had the biggest number of items that had good facility indices (i.e., there were six items with IF between 0.37 and 0.66) and discriminative power (between 0.44 and 0.70). However, three items were outside the expected range of point biserial correlation index (below 0.30). As the majority of the items in this subtest had good item parameter values, the committee decided to keep it, but revised items 1, 3 and 7(those with low discriminative power).

4.4.1.3 Test reliability. One of the categories of reliability is internal consistency coefficient (the others are alternate-form coefficients and test-retest coefficients) and using Cronbach's alpha coefficient as the internal consistency measure is taken as the "industry standard" (Khalifa & Weir, 2009, p. 148). Internal consistency measure such as Cronbach' alpha coefficient shows to what extent individual items function in a similar manner (Popham, 1990). By the same token, however, having test tasks that include different item types or measure different aspects of an ability (such as careful local reading and careful global reading) might not yield high Cronbach's alpha measures (Jones, 2001). Moreover, Cronbach's alpha reliability measure is likely to produce high estimates when the test is normally distributed, and used in longer tests than in shorter texts (Brown, 2002).

Table 38 Cronbach's alpha estimates of all subtests

	Cronbach's alpha	**Item #**
Subtest 1	0.429	8
Subtest 2	0.244	7
Subtest 3	0.495	8
Subtest 4	0.383	7

The estimates of alpha coefficient computed for all the subtests is given in Table 38. The alpha coefficient values in Table 38 would be considered low if it were a test assessing one type of ability, or a scale focused on one type of behavior (for example, Furr and Bacharach (2008) suggest that values around 0.7 and 0.8 show good reliability). However, for a number of reasons the literature cautions about drawing premature conclusions:

- the alpha coefficient is sensitive to text length (Brown, 2002)
- the number of items on a scale (a test in this case) affect alpha coefficient estimates; for example, having more than 20 items may yield an alpha coefficient value of 0.70 or greater even when the correlation among items is very small (Cortina, 1993)
- high alpha coefficient values may not be expected when different traits, or abilities are tested (Jones, 2001).

Hence, though the alpha coefficient values for the subtests in Table 38 may be considered low for unidimensional (testing one trait/ability) lengthy tests with more than 20 items, in the context of this study it may not be appropriate to expect high coefficient values because the reading test taps on different aspects of reading ability (for example, it assesses skimming to find the gist of a text and reading carefully to understand information that is not explicitly given), rather than testing a unidimensional trait. This was also the case in Cambridge ESOL exams which includes a variety of task based materials and item types, and it has been claimed that, as such, it is not appropriate to expect high alpha coefficient values (Saville, 2003). Therefore, the low alpha coefficient values do not cast doubt on the internal consistency of the reading test *per se*.

In Table 39, item-totals statistics are presented; that is, the correlation between the scores on each individual item and the score on the test is computed. In the light of the information given above on the limitations of calculating alpha coefficients in short tests tapping on different abilities, it is not surprising that the Cronbach's alpha coefficient values obtained as a result of item-total correlation computations are lower than the range specified in the testing literature (generally, 0.7 – 0.8 is considered good correlations) (Bachman, 2004). In Table 39, the column that gives the most significant information on the test items is the fourth one titled "Cronbach's Alpha if Item Deleted". This column shows the Cronbach's alpha coefficient value if the individual item is removed from the test. For example, for Subtest 1, the alpha coefficient is 4.29 (Table 38). If item 8 were deleted from the test, the alpha coefficient would increase to 0.47 (Table 39). Therefore, this table could be helpful in deciding which items in a test should be deleted to increase the internal consistency of a test. In the case of the subtests used in this present study, rather than using this information the committee and I have chosen to consider parameters obtained from item analyses, and the degree of correspondence of aspects of each reading text and the items with the test specs.

Table 39 Item – Total statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
		Subtest 1		
Item 1	2.23	2.30	0.12	0.42
Item 2	2.67	1.88	0.36	0.31
Item 3	2.76	2.02	0.29	0.35
Item 4	2.90	2.29	0.18	0.40
Item 5	2.77	2.04	0.29	0.35
Item 6	2.72	2.06	0.24	0.37
Item 7	2.60	2.30	0.03	0.47
Item 8	2.61	2.32	0.02	0.47
		Subtest 2		
Item 1	2.92	1.62	0.17	0.17
Item 2	2.77	1.51	0.23	0.12
Item 3	2.92	1.68	0.11	0.21
Item 4	2.49	1.59	0.25	0.12
Item 5	2.89	1.69	0.10	0.22
Item 6	2.81	1.90	-0.08	0.34
Item 7	2.68	1.85	-0.04	.310
		Subtest 3		
Item 1	5.07	2.14	0.22	0.47
Item 2	5.20	1.78	0.42	0.38
Item 3	5.12	1.90	0.39	0.40
Item 4	5.11	1.95	0.35	0.42
Item 5	5.40	2.09	0.08	0.53
Item 6	5.08	2.17	0.17	0.48
Item 7	5.45	1.96	0.17	0.49
Item 8	5.17	2.16	0.10	0.51
		Subtest 4		
Item 1	3.59	1.91	0.26	0.30
Item 2	3.45	1.96	0.20	0.33
Item 3	3.30	2.11	0.12	0.37
Item 4	3.31	2.01	0.19	0.34
Item 5	3.38	2.11	0.10	0.39
Item 6	3.35	1.96	0.22	0.32
Item 7	3.37	2.11	0.10	0.39

4.4.2 Reading Test V2 results. The second version of the reading test was administered as a whole to 27 participants between April and June 2017. One participant left the study stating that he did not feel well enough to complete the test;

therefore, the following results are from 26 participants who took the reading test and carried out a think aloud procedure while doing so. The primary aim of administering this version to a limited number of participants was to collect qualitative data through verbal protocols in order to examine the cognitive processes of the participants while responding to reading tasks. However, some descriptive statistics and item analyses were also carried out to obtain information on item parameters. The participant numbers and their level groups are given in Table 40.

The participants for the second version of the reading test were chosen from among the advanced, upper-intermediate and intermediate levels because at the time of data collection the students in the pre-intermediate and pilot pre-intermediate groups were to receive another two to three and a half months of instruction before they were allowed to take the proficiency exam. As the test was aimed at a higher level of proficiency than those students were at that time, those groups were not included in the study.

Table 40 Participant numbers and groups

Level Group	**n**
Advanced	9
Upper-Intermediate	7
Intermediate	10

4.4.2.1 Descriptive statistics. In the second version of the reading test, there were four reading tasks with a total of 30 items. Each item had 1-point weight. As can be seen in Table 41, the mean score of the test was 66/100. The minimum score was 11 and the maximum score was 28.

Table 41 Descriptive statistics of Reading Test V2

	n=26
Mean	19.8 (66%)
Std. Deviation	4.70
Variance	22.10
Skewness	-0.045
Kurtosis	-0.716
Range	17.00
Minimum	11.00
Maximum	28.00

Table 42 presents frequency distribution of the scores on a 30-point scale. There is a wide distribution with only two scores (15 and 17) being seen three times, and all the other scores had a frequency of at most two.

Table 42 Frequency of scores

Reading Test V2			
Score	**Freq.**	**Percent**	**Cumulative Percent**
11.00	1	3.8	3.8
12.00	1	3.8	7.7
13.00	1	3.8	11.5
15.00	3	11.5	23.1
17.00	3	11.5	34.6
18.00	2	7.7	42.3
19.00	1	3.8	46.2
20.00	2	7.7	53.8
21.00	2	7.7	61.5
22.00	2	7.7	69.2
23.00	2	7.7	76.9
24.00	2	7.7	84.6
25.00	1	3.8	88.5
26.00	1	3.8	92.3
28.00	2	7.7	100.0
Total	26	100.0	

A final analysis on the test score was the test of normality: it reveals whether the scores are distributed normally within the population. As can be seen in Table 43, the distribution of scores was normal (p=.000).

Table 43 Test of normality

	Shapiro-Wilk		
	Statistic	df	Sig.
Test score	0.557	26	0.000

4.4.2.2 Item analyses: Item facility and item discrimination (V2). The analyses of the items of the second version of the reading test in accordance with CTT conventions are given in this section. Table 43 presents item facility indices (IF), item discrimination indices (d) and point biserial correlations (r_{pbi}) for the items in the test.

The second version of the reading test contained 22 selected response items (multiple choice and matching) and 8 constructed response items (short answer) (see Section 3.3.2.3.1.1 for the content of Reading Test V2).

In terms of item facility, the lowest value (the most difficult) was 0.31 and the highest (the easiest) was 0.96. There were six items that had values higher than 0.80 (Items 2, 5, 12, 13, 19 and 22), and there were four items that had values lower than 0.40 (items 8, 11, 21 and 25). All the ‘easy' items were selected response type items (two matching and four multiple choice items). Three of the ‘difficult' items were selected response (multiple choice) and one was a short answer item.

The items with the lowest facility values (items 8 and 21) were inferencing questions (i.e., participants had to understand implicit information). The item that had the highest facility value was item 19, which required detailed reading of a portion of the paragraph (macro level comprehension).

Table 44 Item analysis of Reading Test V2

Item #	**Item type**	**Item Facility (IF)**	**Discrimination Index (*d*)**	**Point-biserial correlations (r_{pbi})**
1	Matching	0.73	0.50	0.45
2	Matching	0.96	0.13	0.20
3	Matching	0.73	0.63	0.54
4	Matching	0.65	0.50	0.34
5	Matching	0.88	0.13	0.11
6	Matching	0.77	0.25	0.28
7	MC	0.92	0.25	0.42
8	MC	0.31	0.50	0.32
9	MC	0.58	0.13	0.29
10	MC	0.73	0.38	0.38
11	MC	0.35	0.13	0.12
12	MC	0.81	0.13	0.02
13	MC	0.85	-	0.09
14	MC	0.62	0.13	0.15
15	MC	0.62	0.13	0.21
16	MC	0.50	0.38	0.34
17	MC	0.65	0.38	0.22
18	MC	0.69	0.63	0.50
19	MC	0.96	0.13	0.37
20	MC	0.77	-	(0.12)
21	MC	0.31	0.13	0.26
22	MC	0.85	0.25	0.25
23	Short answer	0.50	1.00	0.72
24	Short answer	0.69	0.63	0.61
25	Short answer	0.38	0.25	0.31
26	Short answer	0.46	0.63	0.60
27	Short answer	0.62	0.88	0.79
28	Short answer	0.73	0.63	0.58
29	Short answer	0.46	0.38	0.39
30	Short answer	0.69	0.63	0.48

In terms of the discrimination indices, the lowest index was 0, and the highest was 1. Out of 30 items, 10 had low discrimination indices (between 0 and 0.13) (items 2, 5, 9, 11, 12, 13, 14, 15, 19, 20 and 21), and they were all selected response items. The highest discrimination index belonged to a constructed response item (item 23). The short answer type items had the highest discrimination indices compared to the other types of questions.

As for the point-biserial correlations, the lowest value was -0.21 (item 20, multiple choice), and the highest was 0.79 (item 27, short answer). Out of the 30 items, 13 of them were below the expected value of 0.30 (two items, 6 and 9 had values very close to the threshold, they were 0.28 and 0.29, respectively). Those items that did not discriminate well usually did not have a good correlation coefficient.

Overall, the short answer items had more favorable qualities than the MC items. Items 23-30 (except for item 25 which had a value of 0.38 instead of 0.40), had facility values within the accepted range. Those items also had the highest discrimination indices and the highest correlation coefficients among all the items, which shows that those items reliably measured the intended test construct.

4.4.2.3 Test reliability. For the second version of the reading test, internal consistency coefficients as a measure of reliability were calculated again. The fact that the test was administered as a whole provided the opportunity to calculate the coefficients of items that tap onto specific reading types separately; for example, matching items require comprehension of the main ideas (global careful reading), and short answer items require search reading (global expeditious reading to locate the parts to be read, and then careful reading to extract the necessary information). This method of calculating alpha coefficients separately for each item type resonates with the assumption that the items computed measure the same underlying construct. Khalifa and Weir (2009) recommend that “Where a test consists of items or groups of items which are intended to test different things, then they should never be analyzed together when estimating internal consistency” (p. 149).

The items in Reading Test V2 can naturally be grouped according to the item types used in the test as follows:

1. items 1 – 6 : test careful global reading (matching)
2. items 7 – 22: test various reading skills (careful local and global, expeditious local) (MCI)
3. items 23 – 30: test search reading (short answer).

The reading operations required to answer the first and third group of items is unambiguous: group 1 items are all matching items that require test takers to understand the main idea of a paragraph; therefore, they tap on careful global reading ability, and group 3 items are all search reading items in the short answer format, which require the test takers to search for the location of the relevant answer and then extract the answer from the text However, group 2 items call for the use of different operations: some require local reading whereas others global reading, and therefore, it may not be meaningful to expect high alpha coefficient correlation estimates for those items.

Table 45 Cronbach's alpha coefficient values for Reading Test V2

	Cronbach's alpha	N of Items
Matching (careful global)	0.67	6
MCI (both careful/expeditious and global/local)	0.27	16
Short answer (search)	0.79	8

Table 45 gives the alpha correlation coefficients in total for the three group of items from Reading Test V2. As expected, correlations of the matching and search reading items reveal high reliability (α=0.67 and α=0.79, respectively) whereas the MC items reveal lower reliability values (α=0.27) due to the fact that different aspects of reading ability are tested through those items.

Table 46 Item - total statistics (V2)

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Matching items - Careful global reading				
Item 1	4.00	1.36	0.607	0.543
Item 2	3.77	2.02	0.254	0.672
Item 3	4.00	1.52	0.430	0.617
Item 4	4.08	1.43	0.461	0.606
Item 5	3.85	1.73	0.423	0.625
Item 6	3.96	1.71	0.268	0.676
Short answer – Search reading				
Item 23	4.04	4.51	0.646	0.739
Item 24	3.85	4.93	0.488	0.766
Item 25	4.15	5.33	0.260	0.801
Item 26	4.08	4.55	0.630	0.742
Item 27	3.92	4.55	0.651	0.739
Item 28	3.81	4.96	0.502	0.764
Item 29	4.08	5.19	0.313	0.794
Item 30	3.85	4.93	0.488	0.766

Table 46 reveals item-total statistics: only group 1 and group 3 items are presented here for reasons given above. The ‘Cronbach's Alpha if Item Deleted’ column gives coefficient estimates if that individual item is deleted from the test. According to the table, if item 2 is deleted, the coefficient estimate will increase to 0.672; none of the other items’ deletion increases the coefficient estimates. The reason for this is apparent in Table 43. As can be seen, item 2 has a very high item facility value (IF=0.96), which is probably the reason why its deletion was estimated to increase the internal consistency of the first group of items. Similarly, in group 3 items, the highest value in the ‘Cronbach's Alpha if Item Deleted’ column belongs to item 25. According to Table 43, item analyses results, item 25 was the only item that had a facility value lower than anticipated (IF=0.38). Hence, the internal consistency measures inform that removing that item would increase reliability measures to 0.801.

4.4.3 Summary of the results of item analyses (V2). Descriptive statistics on the first version of the reading test yielded the following mean scores:

Table 47 Subtest averages

Reading test V1	Average (%)
Subtest 1	38
Subtest 2	47
Subtest 3	74
Subtest 4	57
Reading test V2	66

Alderson et al. (1995) maintain that items closer to the facility value of 0.5 should be sought to achieve the widest scope of variation among test takers. Very easy items do not provide any information on the test takers as they are answered correctly by a majority of the test takers. Therefore, the difference between weak and strong test takers cannot be observed. Difficult items are more acceptable if they discriminate between test takers with different ability levels (Khalifa and Weir, 2009). In Reading Test V2, six of the items had facility values above 0.80 (easier than anticipated), and four of them had facility values lower than 0.40 (more difficult than anticipated). The items that were too easy or difficult usually did not differentiate well between participants at different proficiency levels. However, this is accepted as a quality of a criterion-referenced test rather than a weakness *per se* (for a discussion of a criterion-referenced test, refer to Section 5.3).

In V2 of the reading test, the open-ended items were the most successful in terms of their facility value, discriminative power and reliability. As the open-ended items required the participants to produce the answer themselves, they could not use test taking strategies such as elimination of options, or getting clues from the options. Without any options for guessing, those participants who understood the question and managed to locate the relevant section of the text, read carefully and wrote down the answer.

As the CTT analysis revealed, more reliable results on the discriminative power and the correlation between items and total score (point biserial) can be obtained by administering the test to a larger number of test takers. Nevo (1980) recommends around 100 subjects for the ordering of items according to their difficulty levels. He claims,

> The larger the sample, the smaller the standard error of the items' characteristics. The index of difficulty of an item in the population measured by the percentage of correct responses (P), the item-total score correlation in the population (e), and other items' parameters can be estimated more accurately when a larger sample is employed (p. 328).

Despite this information, the number of people recruited to administer the second version of the test was limited with 27 participants. This decision is closely related to the nature of the research question RQ2b. RQ2b focused on the cognitive processes that are activated during test taking, and the data collection method was set as think aloud protocols. Due to the shortcomings of dealing with – collecting, transcribing, coding, etc. – recorded media, a limited number of participants could be recruited. The nature of the research question necessitated a thick description of cognitive processes, rather than statistical information. Consequently, it may not be possible to assert, with confidence, that the reliability coefficients of some items of the Reading Test V2 are appropriate. However, further research might shed light on the statistical properties of this version of the test.

4.5 Summary of findings of Research Question 1, 2 and 3

Table 48 Summary of findings of Research Question 1, 2 and 3

<table>
<tr><th colspan="3">RQ1 – Context validity</th></tr>
<tr><td><b>Reading test construct definition</b></td><td colspan="2">The cognitive processing model of reading emphasizing the use of both bottom-up and top-down reading models interactively</td></tr>
<tr><td><b>Criterial parameters</b></td><td colspan="2">The literature on academic reading needs and the findings of the needs analysis study guided the decisions on the parameters of test facets provided in the sociocognitive framework.</td></tr>
<tr><td><b>Test specifications</b></td><td colspan="2">Assessment of reading was based on two reading types: careful reading and expeditious reading. Both local and especially global reading was emphasized.</td></tr>
<tr><th colspan="3">RQ2 – Cognitive validity</th></tr>
<tr><th colspan="3">Retrospective Investigation</th></tr>
<tr><td><b>Preview strategies</b></td><td colspan="2">No distinctive pattern overall. GR1 participants used reading quickly and selectively more frequently than GR2 participants.</td></tr>
<tr><td><b>Test response strategies</b></td><td colspan="2">Reading slowly and carefully, re-reading a part of the text and search and match similar</td></tr>
<tr><td><b>Location of the answer</b></td><td colspan="2">Across sentences</td></tr>
<tr><th colspan="3">Introspective Investigation</th></tr>
<tr><th>Item Type</th><th>Expected Reading</th><th>Findings</th></tr>
<tr><td>Vocabulary</td><td>Careful local reading</td><td>Careful reading.
Reading sentences adjacent to that with the vocabulary item.
Mainly global reading.</td></tr>
<tr><td>Macro-level comprehension</td><td>Careful global & local reading</td><td>Careful reading.
Scanning.
Test management strategies: elimination of options.
Building a mental model of the text.</td></tr>
<tr><td>Matching</td><td>Expeditious global reading</td><td>Careful reading.
Establishing a mental representation of the text.
Monitoring.
Meaning based selection of options.</td></tr>
<tr><td>Search Reading</td><td>Expeditious global & Careful global reading</td><td>Careful reading.
Building a mental model of the text.
Test management strategies: selecting options through meaning.</td></tr>
</table>

Table 48 (Continued)	
RQ3 – Scoring validity	
Reading Test V1	Score range: 38% - 74% IF indices: 15 items within the range of 0.40 – 0.80. 9 Items *too difficult*, 6 items *too easy*. 27 Items discriminated well. 23 items showed good item-total score correlations. Cronbach's alpha range: 0.24 – 0.50
Reading Test V2	Average score: 66% IF indices: 19 items within the range of 0.40 – 0.80. 4 Items *too difficult*, 7 items *too easy*. 19 Items discriminated well. 17 items showed good item-total score correlations. Cronbach's alpha range: Matching: 0.67 Multiple choice: 0.27 (careful & global reading items computed together) Short Answer: 0.79

CHAPTER 5

DISCUSSION AND CONCLUSION

5.1 Introduction

Validity is considered to be a primary concern in all testing situations (Bachman, 2000, 2005, Bachman & Palmer, 1996, 2013; Chapelle, Jamieson, & Hegelheimer, 2003; Fox, 2004; Kane, 2012; Lazaraton, 2002; Lissitz, 2009; McNamara, 2006; Milanovich, Saville, Pollitt, & Cook, 1996; Sireci, 2016; van der Walt & Steyn, 2008). However, the definition of validity went through a major transformation within a few decades: from being considered a characteristic of a test, it came to be accepted as the extent of justification one could provide for the score interpretations. Since then, various approaches for test validation, and specifically frameworks on which to build validation methods have been proposed (Kane, 2016; O'Sullivan & Weir, 2011; Weir, 2005a). Of the many, Weir's (2005) socio-cognitive framework offers a clear outline in the planning and sequencing of validation work (Taylor, 2014). Hence, this study utilized a reading model (Urquhart & Weir, 1998; Weir & Khalifa, 2008a) and the socio-cognitive framework (Weir, 2005) for the validation of a reading test. The sociocognitive framework comprises five aspects of validity; namely, context, cognitive, scoring, consequential and predictive validity. Test taker characteristics are also included as a component of the framework. Weir (2005) posits that there is a reciprocal relationship between the components of the framework. Khalifa and Weir (2009) assert that context validity, cognitive validity and scoring validity constitute *construct validity*, an overarching concept that incorporates all types of validity evidence (Messick, 1995). Following this approach, three types of validity evidence were generated through the investigation of the following research questions:

1. Contextual validity was investigated through Research Question 1: How is academic reading ability conceptualized and operationalized as a test construct?
2. Cognitive (theory-based) validity was investigated through Research Question 2: What are the cognitive processes that underlie the construct of the reading test in retrospection and in introspection?
3. Scoring validity was investigated through Research Question 3: To what extent do item parameters contribute to the validity claims of the test?

In order to answer the research questions, first, reading ability was defined as a test construct, and its specifications were drawn in accordance with the academic reading requirements identified in the literature and in a school report (See 4.2.1 and 4.2.2). This is considered the first stage of *a priori* test validation, and the outcome is a test specifications document that presents the criterial parameters of the reading ability within this specific context.

The second stage of *a priori* test validation is the examination of the cognitive validity of the test, which answers the second research question, through retrospective and introspective verbal protocols. The participants' reports revealed the extent of congruence between the processes activated through the test items and the processes that were hypothesized to represent reading activity in real life academic contexts.

Finally, the third research question corresponds to the *a posteriori* validation stage of the socio-cognitive framework, and it investigated item difficulty, item reliability and item discrimination to reveal whether the statistical values derived from the scores corresponded with the expected values as specified in the literature in terms of the difficulty level, the discrimination power and the reliability of the items.

In this chapter, first the results of the study are individually discussed and compared and contrasted with the findings of similar studies in the field. Then, the implications and limitations of the current study, and recommendations for further research are presented.

5.2 Discussion of Research Question 1

The first research question aimed to generate evidence on the context validity of the test by defining contextual parameters using a theoretical framework. Khalifa and Weir (2009) claim that if we can accurately describe the criterial parameters of the reading activities carried out in the target language domain, and operationalize them appropriately then the test takers' performances can be generalized beyond the testing situation.

Context validity, one of the six aspects of validity (see Figure 4 for the validity scheme) in Weir's (2005) framework, is about "the appropriateness of both the linguistic and content demands of the text to be processed, and the features of the task setting that impact on task completion" (Khalifa & Weir, 2009, p. 81). This statement echoes Messick (1995) whose work on validity theory had a significant impact on approaches towards generating validity evidence in educational assessment.

According to Messick (1995), *content relevance* and *representativeness* are two important issues related to the content aspect of construct validity. *Content relevance* is achieved by "determining the knowledge, skills, attitudes, motives, and other attributes to be revealed by the assessment tasks" (p. 12) through analysis of tasks, curriculum, and the nature of the domain processes. *Representativeness,* on the other hand, is about selecting tasks that are functionally important in the target language use domain. Brunswik (1956) calls this *ecological sampling,* which refers to sampling in such a way that all important parts of the construct domain are covered.

In brief, in test construction, *content relevance* and *content representativeness* are achieved through the analysis of the test domain. The description of that domain produces a test specification document that addresses the content aspect of construct validity, or as referred to Weir's framework, context validity. The test specifications document is an important declaration for the relevance and representativeness of the test tasks with relation to the content domain. Hence, constructing test tasks according to this test specifications document provides evidence for contextual validity of the test.

In the present study, *content relevance* was achieved through the analysis of communicative tasks in the target language use domain, specifically those in the first year of academic studies in all five faculties at METU (the engineering, architecture, education, economic and administrative sciences, and arts and sciences faculties), and the criterial levels of achievement in reading. *Representativeness*, on the other hand, was achieved through a careful selection of tasks that were indicated to be important by the stakeholders such as the faculty members, and the instructors of the DBE and MLD.

The literature warns us about the two major threats to validity with relation to the content of the test:

> One is construct underrepresentation – that is, the test is too narrow and fails to include important dimensions or facets of the construct; the other is construct-irrelevant variance – that is, the test is too broad and contains excess reliable variance associated with other distinct constructs, as well as method variance making items or tasks easier or harder for some respondents in a manner irrelevant to the interpreted construct (Messick, 1992, p. 1491).

Construct underrepresentation happens when the content does not sample the tasks in the target domain adequately, or important aspects of the construct are not captured in the test tasks, in which case the assessment becomes invalid (Kaplan & Saccuzzo, 1982). Having trivial content, or too few exam items may lead to construct underrepresentation. In order to avoid the pitfall of construct underrepresentation, utmost care has been taken to make sure the test specifications document for the current test adequately reflects the test construct.

The second major threat, *construct-irrelevant variance*, can either be construct-irrelevant difficulty or construct-irrelevant easiness. In the former, a feature of a task that is unrelated to the construct of the test may cause a group of test takers to score low, leading to bias in scoring. It is also considered unfair test use. The latter happens when some test takers manage to respond correctly using methods irrelevant to the test construct, such as being familiar to a reading text in a reading task. The result of construct-irrelevant easiness is that those test takers with familiarity to a reading text receive scores higher than they would. Other sources of construct-irrelevant variance can be guessing, using test-wiseness strategies, or having poorly constructed items that make it difficult for test takers to understand the gist of the question. In order to ensure that there is no construct-irrelevant variance in the reading test, the psychometric qualities of the items were closely examined, and those items that revealed too high or too low item facility values were re-examined, with some being revised while others were replaced.

Fulcher (2010) mentions other critical advantages of having a test specifications document: it allows the test writers to develop a test for a particular administration, but also ensures the design of parallel forms for each administration of the test.

In a similar vein, Bachman and Palmer (1996) list four purposes for test blueprint which includes specifications for each type of task:

1) to permit the development of other tests or parallel forms of the test with the same characteristics;

2) to evaluate the intentions of the test developers;

3) to evaluate the correspondence between the test as developed and the blueprints from which it was developed;

4) to evaluate the authenticity of the test (p. 177).

In the present study, test specifications were drawn with a perspective similar to that in the literature. The three main purposes specified for test specifications were:

a) to provide guidance to test writers to produce items/tests with similar characteristics;

b) to provide a guideline to the administrators to evaluate whether the developed test form corresponds to the intended operations;

c) to keep a record of the modifications on the test form and the reasons for it.

While generating test specifications is beneficial for the stakeholders of the test, there are some key characteristics which make it valuable and meaningful. These were mentioned by Davidson (2012) as being

- generative: the spec is intended to produce many equivalent test items/tasks;
- iterative: the spec evolves over time and proceeds through versions;
- consensus-based: the spec is co-authored by a team (p. 201).

In developing the specs for the reading test, those three characteristics were taken into account. The specs were detailed enough to allow for the developing of items that consistently test what is intended to be tested and it was made clear that the specs document should be reviewed and revised when necessary. Although I designed the document, the criterial parameters were set during the meetings with the testing committee, hence, consensus on all parts of the document was achieved before it was presented to the administration.

Finally, a crucial aspect of the test specifications document presented in this study is the cognitive processing reading model by Urquhart and Weir (1998), and Khalifa and Weir (2009), which specifies the theoretical underpinnings of the reading test construct. The contribution of the theoretical model to the specification of the construct is significant: the test construct is delineated by the model, and the test specifications reflect facets of the model (Unaldi, 2004); that is, the reading construct is limited to, and defined within the boundaries of the cognitive processing approach (for a detailed discussion of the theoretical model of reading see Section 4.2.1).

The congruence between the facets of the model and the specifications of the behavioural domain is important. Urquhart and Weir (1998) define reading ability on two dimensions: local/global and careful/expeditious reading. They suggest that the purpose of reading determines the type of reading that is going to be carried out. They also assume an interactionist view of reading. That is, while reading readers use bottom-up processing (get information from the text), and top-down processing (use their own knowledge about the topic/genre/discourse); while reading they also elaborate on what they read and monitor their understanding of the text. Apparently, a number of processes at various levels of attention take place during reading, and it may not be realistic to categorize those processes as distinct from each other as has been delineated in the reading model by Khalifa and Weir (2009). Moreover, the reading processes observed in retrospective and introspective protocols suggest a combination of the use of reading types rather than a compartmental reading activity as described in the model. Therefore, the test specifications document was designed in such a way that it reflects these dimensions and processes as it relates to the theoretical underpinning of the model.

5.2.1 General outline for test specifications. The literature and local reports guided the decisions about various aspects of the reading test; however, the testing committee's views, school regulations and practical considerations were also taken into account in reaching the conclusive decisions about the contextual features of the test.

The test specifications document starts with the general purpose of the test and a brief reference was made to the target language use tasks that had been specified in the needs analysis document. The school report helped to identify the broad skills to carry out the reading requirements in academic programs. Those skills were articulated using the Council of Europe's (2009) classification for ease of reference for the future, as setting a level for the test in accordance with the Council's scaling approach is being planned as further research.

A brief description of the test-takers profile was included in the specifications document. Weir (2005) states that test taker profile is closely related to the cognitive validity of the test because test takers' characteristics influence the way they process the test tasks. Moreover, it is important to contemplate on these characteristics to prevent any group bias; that is, the test should not systematically disadvantage a certain group of test takers. For this reason, test developers should consider test takers in choosing tests with appropriate content, and be aware of bias against or for test takers from different backgrounds (Reynolds & Suzuki, 2003).

The specifications document states a test level that was defined using the illustrative global scale from the Common European Framework of Reference (CEFR)(2001). The tentative level set for the test is B2. B2 level *overall reading comprehension* is described using the following statement in the reference book by the European Council (2001):

> Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low frequency idioms (p. 69).

The literature provides information on how to set a CEFR level for an assessment battery: it involves an extensive iterative evaluation process by judges, usually experienced instructors who are familiar with the illustrative scale, and the test taker population. It is beyond the scope of this study to generate evidence for the CEFR level of the test; however, it was found beneficial to set a provisional level to create an awareness of an external criteria for the test, and establish an agreement among item writers about the expected difficulty level of the test tasks.

Nonetheless, as the can-do statement given above for the reading ability at B2 level is too broad a guidance for choosing texts and for the expected operations based on test tasks, various publications were referred to for more information (for example, Davidson & Fulcher, 2007; "Introduction to the CEFR with checklists of descriptors – Eaquals," n.d.; Language Policy Division, 2009; Lowie, Haines, & Jansma, 2010; Martyniuk, 2010; Pearson ELT, n.d.; Takala, 2010).

Definition of the test construct in terms of the expected operations were given in the two tables that presented skills taxonomies for a) careful reading and b) expeditious reading (APPENDIX E).

Careful reading is carried out as either global or local careful reading. In global careful reading, the reader generates a macro-level representation of the text by understanding the main ideas/arguments, and the explicit and implicit propositions. Local reading is operationalized as using contextual clues to understand the meaning of an unknown word.

Three of the test tasks assessed global reading as this type of reading is believed to realistically reflect the reading practices of students in academia. A similar indication was made in a number of studies researching academic needs (such as Enright et al., 2000 and Weir, 1983).

The taxonomy for expeditious reading, on the other hand, included three sub-operations: skimming, scanning and search reading. Although referred to separately in the specifications document, it was decided to test those three operations under one reading task, and in an integrated manner.

Weir (1983) claims that students often need to carry out search reading to obtain information on a specific topic for their assignments. He particularly specifies skimming and extracting important information from texts as a major reader purpose in academia, which shows that the search reading operation is a relevant addition to the new reading test.

The test specifications document is not a fixed and ultimate guide to a test; it should be flexible to reflect any changes that are deemed necessary to be able to justify the score-based interpretations. For example, any changes in the test taker profile, or contextual parameters, such as the communicative functions expected of the test takers in the target language use domain ought to be reflected on the test tasks so that one can justify the value of the decisions made about the test takers' ability in the relevant domain. The version of the test specifications document presented in this study is also subject to change when and if the testing committee decides to improve an aspect of the test to increase its reliability as a tool to predict university students' future performance in academic programs.

Developing a test specs document – a blueprint – with sound theoretical basis and consistent with the communicative needs in the target language use domain helps to achieve both situational authenticity (that the test takes into account the contextual requirements of the tasks) and interactional authenticity (that the cognitive activities of the test taker in performing the test task is similar to that in performing tasks in real life) in a test (Bachman & Palmer, 1996; Douglas, 2000). The term *authenticity* in the works of previous researchers translates into Weir's framework as validity. Though there are differences in terminology, in a broad spectrum of methodological approaches, validity inquiry encompasses gathering both theoretical and empirical evidence as does the present study (see for example, Haertel (1985) and Chapelle (1998)). As such, the design and implementation of the test specifications document presented here verifies content-relevance and representativeness claims of the construct theory, and generates evidence of validity. This evidence justifies the meanings ascribed to test scores, which is congruent with the approach to validity acknowledged in this study: "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (Messick, 1989b).

The approach presented here in the design and development of test specifications have significant contributions to the field: First, at the conceptual level, it explicitly demonstrates the relation between observed performances and the traits relevant to the test. In other words, the connection between theory and data is clearly presented. This undertaking takes test development procedure a step forward for all stakeholders in any assessment situation. Similar to Jamieson (2013), the present study asserts that explicating and advancing the meaning of a construct in language assessment ensures that the stakeholders learn "to look beyond the term [construct] itself, and examine the nuances of its use". A deep and thorough understanding of the concept and being able to position it within the context of use will guide the development of assessment instruments consistent with the needs, and the theory, and will thus address the issues of reliability, validity and fairness.

Secondly, this endeavour is another step in establishing test design and development standards within the domain of second language testing. Especially in high stakes testing, concerns over the appropriateness of the interpretations based on test scores on the part of all stakeholders drive the related parties to follow procedures such as those established in the present study. Burgeoning of sound, theory driven test design and development activities will help to establish certain standards in the realm of language testing, such as the acknowledgement of the responsibility of institutions to generate and present validity evidence to their stakeholders.

Thirdly, validation efforts inform language teaching practices. The association between language assessment and language instruction is immediate. Instructional curriculum is one of the factors affecting test design. The content of the assessment instrument is partly derived from the curriculum, syllabus, and other documents used in instruction. In terms of content-relevance, the test developers need to make sure that they sample tasks that can cover aspects of competence so that they can support their claims regarding the generalizability of the scores beyond the testing situation. The better sampling of tasks from the instructional domain, will impact instruction positively as the teachers will shy away from “teaching to the test”.

5.3 Discussion of Research Question 2

The second research question aimed to generate evidence on the cognitive validity of the reading test through retrospective and introspective investigation of cognitive processes that were activated while carrying out tasks in the reading test. To this end, the data collected retrospectively and introspectively were analyzed separately to establish the extent of utilization of the cognitive processes affirmed in the scheme in the reading test, and the correspondence of the reading types and purposes with the new reading paradigm. As has been popularly reiterated in the literature, the match between the construct and the test would demonstrate that the test actually measures what it claims to measure (Kelley, 1927).

The findings of the retrospective and introspective protocol forms demonstrated that the operationalization of reading construct through the test supports the interpretations of test scores; that is, the cognitive processes elicited through the protocol forms were congruent with the theoretical framework of the test, and the reading purposes and dimensions (local vs. global) were consistent with the behavioral criteria defined in the test blueprint. Therefore, the reading test (V2) was found to be suitable as a measurement tool with regard to the cognitive processes hypothesized to constitute reading ability. Furthermore, the data supported the argument that reading ability is componential (Devi, 2010; Urquhart & Weir, 1998), and therefore, the reading model is an appropriate tool in the design of the reading test.

5.3.1 Retrospective findings. Retrospective data was collected through a protocol form during the administration of the first version of the reading test in four separate subtests. The retrospective form was designed in such a way that the Parts B, C and D corresponded to the relevant sections of the reading model (Goal Setter, Central Processing Core, Knowledge Bases) (see Section 4.2.1 for details on the reading model). The relation between the parts of the retrospective form and the reading model is given in Table 49.

Table 49 Relation between the parts of the protocol form and the reading model

			Goal Setter	Central Processing Core	Knowledge Bases
Part B	1	read whole or part of the text slowly and carefully	careful reading		
	2	read whole or part of the text quickly and selectively to get a general idea	expeditious reading		
	3	did not read the text			
Part C	1	match words in the question with exactly the same words in the text	scanning		
	2	quickly match words in the question with similar or related words in the text	search reading		
	3	look for parts of the text that the writer indicates to be important	global reading		
	4	read key parts of the text such as the introduction and conclusion	global selective		
	5	work out the meaning of a difficult word in the question	local reading		
	6	use knowledge of vocabulary			lexical
	7	use knowledge of grammar			syntactic
	8	read the text or part of it slowly and carefully	careful reading		
	9	read relevant parts of the text again	careful reading		
	10	use knowledge of how texts like this are organized			text structure
	11	connect information from the text with one’s knowledge			general/topic
Part D	1	Found the answer within a single sentence		propositional level	
	2	… by putting information together across sentences		mental model level	
	3	… by understanding how information in the whole text fits together		text level	
	4	Knew the answer without reading the text			general/topic
	5	Could not answer the question			

In populating the table with the frequencies obtained through the retrospective protocol form, only GR2 participants' responses were used with the assumption that those who scored high are most likely to have used the appropriate strategies to find the answers to the questions. When the frequencies were mapped on to the reading model (in the manner as shown in Table 49), it was seen that the test items elicited mainly careful reading strategies at the global and local level (F=9.08) (Table 50). After careful reading, expeditious reading (F=4.77) was the most frequently used strategy. This strategy refers to global selective reading, that is, the reader selects parts of the text such as the introduction or the conclusion to achieve at a text level understanding.

The results revealed that careful reading played an important role in answering questions in Reading Test V1. Findings on the use of careful versus expeditious reading strategies in similar (academic) tests are contradictory: In one study by Weir et al (2006), the majority of the participants (61%) read the text quickly and selectively (expeditious reading at the global level) before reading the questions whereas in another study, Khalifa and Weir (2009) reported that in the B2 level Cambridge ESOL reading exam there was a clear coverage of careful reading at the global level but a limited coverage of expeditious reading at the global level (p. 126).

Table 50 Mapping usage frequencies on relevant sections of the reading model

		Frequency
Reader purpose	Careful Reading – Global & Local	9.08
	Expeditious reading - Skimming	4.77
	Expeditious reading - Scanning	0.92
	Expeditious local reading – Search	2.21
Cognitive processes	Creating a representation of the text(s)	1.94
	Building a mental model	3.79
	Establishing propositional meaning	1.23
Knowledge base	Text structure knowledge	0.60
	General/Topic knowledge	0.79
	Syntactic knowledge	0.63
	Lexical knowledge	1.77

The other two strategies that were used at lower frequencies were search reading (F=2.21) and scanning (F=0.92). Search reading and scanning relied mainly on the lexical access component in the Central Processing Core. Especially participants in GR1 relied more on lexical recognition and word matching rather than using textual features (such as discourse structure, or subtitles) to identify sections relevant to the questions.

The use of careful reading and expeditious reading strategies were not exclusive; that is, they were used in a cohesive manner by the participants in the four subtests. Participants who reported that they used *S2 – Search & match similar,* an expeditious reading strategy, also marked *S8 – Read slowly and carefully*, and *S9 – Re-read parts*, which are careful reading strategies carried out at global and local level.

In terms of cognitive processes, building a mental model (F=3.79) had the highest frequency. This finding is similar to the Weir et al (2006) study, in which 89% of the participants used the strategy *putting information together across sentences*, which refers to 'building a mental model of the text' in our reading model.

Building a mental model refers to processing incoming information and integrating the new information into a mental representation of the text created so far. Field (2004) states that

> Incoming information has to be related to what has gone before, so as to ensure that it contributes to the developing representation of the text in a way that is consistent, meaningful and relevant (p.241).

By doing so, the reader identifies the main ideas, relates them to the information previously read, makes a distinction between important ideas and supporting details, and thus, builds a macro structure of the text as more is read. Comprehension in this hierarchical manner is congruent with the cognitive model which was also observed in the analysis of the introspective data.

Following mental model level, in the second place was text level representation (F=1.94). This process requires the ability to "recognize the hierarchical structure of the whole text and determine which items of information are central to the meaning of the text" (Khalifa & Weir, 2009, p. 53). Considering the degree of knowledge – of discourse structure, coherence between ideas, etc. – it is not surprising that the frequency rate is low, and predominantly the high-scoring participants reported to have created a representation of the text during the test.

In terms of knowledge base, lexical knowledge was used the most (F=1.77) and knowledge of text structure (F=0.60) was used the least. This finding shows that the participants used local reading strategies more frequently than global reading strategies. In the use of lexical knowledge, the difference between GR1 and GR2 participants was significant. Having a wide range of lexical knowledge helps text comprehension; together with grammatical knowledge it also affects reading test performance (Kobayashi, 2002; Shiotsu & Weir, 2007).

In brief, in terms of reading types, careful reading, mostly global and occasionally local, was the primary reading style employed by the participants. This finding suggests that the macro comprehension and critical items were responded as the test writers intended. The specific items that were expected to initiate expeditious reading strategies failed to do so; hence, in the second version of the reading test, different type of questions (search reading in the short answer format) were included to address this deficiency. Local expeditious reading was used more frequently by those participants in GR1. They relied, more than others, on word recognition skills to identify the location of the text relevant to the question, or to select an option that they thought correctly answered the question.

In terms of cognitive operations, both bottom-up and top-down processes were activated while answering the questions: GR2 participants, i.e. the advanced level at the DBE, used top-down processes more often than the others: e.g. they used their knowledge of discourse and text organization more frequently than did GR1 participants.

5.3.2 Introspective findings. The second set of data were obtained through verbal protocols. Verbal reports have been widely used in many fields including reading research, and especially in the analysis of cognitive processes (see, for example, Alison Green, 1998; Kucan & Beck, 1997; Pressley & Afflerbach, 1995). In the present study, the verbal reports obtained from the participants were coded using a rubric adapted from Cohen and Upton (2006). The data from the verbal protocols were primarily analyzed qualitatively, with the aim of understanding the strategies used during test taking as part of the process of examining the cognitive validity aspect of the test. The analysis revealed important findings about the cognitive processes as well as the relevance of the reading model for academic reading in the local context.

The first finding is that the participants approached the texts as a test-taking task. In the context of this study, academic reading is mainly characterized as reading for information and argument (see Section 4.2.2 for the academic reading needs); that is, reading to learn. However, the participants' main aim in dealing with the texts was correctly answering the test questions. This finding is similar to that of Cohen and Upton (2006) who investigated test takers' cognitive processes during a TOEFL test of reading. They reported that the main goal of their participants was to find the right answer to the questions rather than learning from the texts they read.

A second finding is that during test taking a wide variety of strategies were used by the participants regardless of their level of proficiency. It was generally a combination of strategies that allowed them to answer the questions on the reading test. Anderson (1991) reports a similar finding saying that "there is no single set of processing strategies that significantly contributes to success [...]. Readers scoring high and those scoring low appear to be using the same kind of strategies while reading and answering the comprehension questions [...]." (p.468).

In this study, reading is defined as an interactive process, during which the reader makes decisions as to how to read (goal setting), activates processes (bottom-up and top-down) and uses her knowledge base (lexical, syntactic, world, etc.) to decipher the meaning in the printed text. As such, it requires many processes to be activated during reading. In test taking too, many processes, and knowledge sources were activated to complete the tasks as needed. Therefore, though mainly approached as a test task, the reading test elicits processes similar to those in real life reading. This finding is evident in Table 51 which shows the types of processes activated during test taking by both the high-scoring and low-scoring participants.

Table 51 Relation between test taking processes and the reading model

		Low scorers		High scorers	
		#	FR	#	FR
	Goal Checking	41	4.5	26	2.8
Reader purpose	Careful Reading - Global	281	31.2	270	30.0
	Careful Reading – Local	57	6.3	74	8.2
	Expeditious reading - Skimming	20	2.2	17	1.9
	Expeditious reading - Scanning	96	10.7	86	9.6
	Expeditious reading - Search	23	2.6	24	2.7
Cognitive processes	Creating a representation of the text(s)	18	2	21	2.3
	Building a mental model	41	4.6	64	7.1
	Inferencing	67	7.4	84	9.3
	Establishing prepositional meaning	15	1.7	13	1.4
	Syntactic parsing	8	0.9	7	0.8
	Lexical access	24	2.7	33	3.7
Knowledge basis	Text structure knowledge - Genre	11	1.22	9	1.00
	Text structure knowledge – Rhetorical tasks	3	0.33	4	0.44
	General knowledge of the world	-	-	-	-
	Topic knowledge	-	-	1	0.11
	Meaning representation of text so far	-	-	3	0.33
	Syntactic knowledge	-	-	-	-
	Lexicon	-	-	-	-

In Table 51, the results are presented for two groups of participants: according to their scores on the test, the participants were divided into three groups, and the top and bottom groups are represented on the table as those who scored low (N=9) and those who scored high (N=9). The amount of use of each strategy (#) was mapped onto the reading model and frequency rate of cognitive processes per person (FR) are given as well. The frequency rate was calculated in a more detailed manner for this data set: the number of times the participants mentioned using a certain strategy was classified according to the question type it was used for (for example, careful reading, or matching). Then, the frequencies were divided by the number of items of that type. This approach allowed to reveal whether the different item types could elicit relevant cognitive processes during test taking.

Comparing the findings of Reading Test V1 and V2, careful reading was the predominant reading style employed during test taking in both versions. Different from V1, in V2, it was followed by scanning, and then search reading and skimming. The difference in the frequency order is probably due to the inclusion of the search reading task in the second version of the test. In the search reading task, there was a long text (about 3000 words) and eight open-ended questions. The participants were expected to find the section relevant to the answer and then read carefully to extract the answer. In locating the relevant section, the participants chose words in the question as keywords and scanned them through the text to find the relevant section. Hence, scanning was the second strategy in the frequency rank list.

In terms of cognitive processes, both groups used a combination of top down and bottom up processes without much difference in the usage ratios, perhaps except for *enriching the proposition*, which is a process related to *building a mental model*. Building a mental model refers to the process of understanding the main ideas of the text, relating them to ideas previously stored in memory, and forming a cohesive link between them so as to build up a macro structure of the text. During this process, the reader monitors her own comprehension to check whether her interpretation of the text is consistent with the meaning representation established so far. Weir and Khalifa maintain that weaker readers may lack this type of monitoring (2008a). The findings presented here support their argument.

Another difference between the low- and high-scoring participants was in inferencing. Inferencing is going beyond what is explicitly stated in the text (Oakhill & Garnham, 1988). It also requires the use of world knowledge in the relevant area (Nuttall, 1996). High scorers more efficiently carried out inferencing because they can deal with the lower level processes such as word recognition, lexical access and syntactic parsing in an automated manner, and therefore, they have more capacity to process information at the higher levels (Khalifa and Weir, 2009).

The participants also used their knowledge base while answering the test questions: *knowledge of genre* and *knowledge of rhetorical structure* were involved during test taking more than others. The use of *general knowledge of the world* was not expressed openly though it is probable that during *building a mental model,* the participants used their world knowledge while judging the coherence and consistency of incoming information while integrating it into the mental model.

Different item and response types affected the participants' use of strategies in the test; therefore, each item type is discussed separately below.

The vocabulary items were designed as careful local reading items (for an example of a vocabulary item see APPENDIX H). The results revealed that both global and local reading (requiring bottom-up and top-down processes) were carried out in answering these items. Careful reading was the major reading strategy used. Rather than inferring the meaning from a single sentence, as anticipated, neighboring sentences were used to resolve lexical ambiguity.

The participants mainly relied on their understanding of the meaning conveyed in the passage when answering these questions. However, they also used an elimination strategy to choose an option about half of the time. It was mainly the low scoring participants who relied on an elimination strategy. This strategy did not, however, warrant a correct response at all times. High scoring participants answered correctly (94%) more than twice as much as low scoring participants, which suggests that it was a challenging item type for low scoring participants.

Rereading the text was one of the most frequently used strategy in answering vocabulary questions, together with repeating, translating and paraphrasing. The literature specifies rereading , among others, as one of the strategies helping comprehension (Grabe & Stoller, 2011).

The reading strategies that were not utilized as much as others in vocabulary items were those that referred to making meaning from the whole of the passage such as *adjusting comprehension of the passage as more is read* (RM15) or *using knowledge of the passage* (RM22).

Comparing the strategies elicited through the vocabulary items and the others, it was seen that the vocabulary items acted similar to macro level comprehension items: in both item types reading a portion of the text carefully, and reading a portion of the text one more time carefully were used the most. The other processes that were activated while answering vocabulary items were not exclusive to this item type. Therefore, it was seen that testing the skill "guessing meaning of unknown words" with what is called 'vocabulary items' did not provide any extra information about the test taker.

Judging the results, the committee decided not to include the item type 'vocabulary' in the test, but inquire other ways to test vocabulary knowledge (such as word form, word meaning, collocations, etc. rather than guessing meaning) in the future.

The matching items were intended to reflect how well the participants can perform expeditious reading in a real life context (for examples of matching items see APPENDIX H). Weir and Khalifa (2008a) define expeditious reading as quick and selective reading to access required information. In expeditious reading, the reader may read at the global (skimming) or local (scanning) level, or conduct search reading which may be both local and global, and the reading direction does not have to be linear.

The matching items in the reading test were intended for expeditious global reading, that is, skimming to get the gist of the paragraphs. It is also known that in expeditious reading reader's knowledge about the structure of the text and background knowledge of the topic can have an important role (Weir et al., 2000). The matching questions were the very first six questions of the reading test. In the case of low scoring participants, lack of knowledge of text structure or the need to activate their schemata before answering the questions may have led them to more detailed reading rather than carrying out the expected skimming strategy.

Comparison of high and low scoring participants' results revealed that they employed processes differently on the reading test. The high scoring participants' verbal reports revealed the use of both expeditious local and careful global reading strategies. The primary difference with the low scoring participants was in the use of keywords in the question to search and match with the keywords in the paragraphs. This strategy proved to be successful for the high scoring participants who, on average, had a correct response rate of 96%. However, the scarce use of skimming as in global expeditious reading was puzzling, as this was expected to be the easiest way to arrive at an answer. Some respondents, after reading the question, rather than making an informed decision about how to proceed, set out to read the options. Inadvertently, option statements contained content words that might have been found to be important as keywords and the respondents scanned them in the text in search for the correct answer.

The majority of the low scoring participants used careful reading strategy – almost three times more than the high scoring participants – in answering the matching questions. Moreover, they almost did not use local expeditious reading strategy at all in their attempt to find the correct answer. The correct response rate for this group of participants was 60%. Weir's (1983) early study on the non-native speakers' difficulties in academic reading also revealed that they struggled with using expeditious reading strategies. Devi's (2010) study reports that non-native speakers of English experience difficulties when they need to conduct quick and selective reading. She also suggests that this variation between different ability students supports Urquhart and Weir's (1998) argument regarding the componentiality of reading, i.e., it is divisible into underlying skills and strategies.

The macro level comprehension items were intended to assess the participants' ability to conduct careful reading at the local and global level (for examples of macro level comprehension items see APPENDIX H). In careful reading, the reader extracts meaning within a sentence up to a paragraph or text level (Weir, 2013). The MALC questions included items focusing on a single sentence (3 items), on a paragraph (10 items) and on the whole text (1 item). Judging by the number of items, careful reading has been emphasized over expeditious and search reading in Reading Test V2 (%53 of the items in total). Though perhaps in different ratios, many high stakes tests predominantly emphasize careful reading skills (see, for example, Katalayi & Sivasubramaniam, 2013; Weir et al., 2009).

The extensive use of this item type is due to its association with academic purpose of reading (A. D. Cohen & Upton, 2006). Academic reading involves using multiple strategies: integrating and connecting information to establish a meaning representation of the text as a coherent whole (Enright et al., 2000), synthesizing information from different parts of a test, or from different texts (Grabe, 2009), knowledge of and ability to use metacognitive strategies to monitor progress in reading (Ellis, 1994; Jun Zhang, 2001), etc. Defined in such a wide spectrum, careful reading for academic studies (i.e. reading for information and argument in the context of the present study) encompasses strategies used in an integrated manner.

The findings reveal the use of a variety of reading and test management strategies in careful reading, which suggests that the test items managed to simulate academic reading processes that take place in real life; however, one point merits consideration. There was only one text-level careful reading question in the reading test. Obviously, the process 'generating a representation of a text as a whole' has not been sampled adequately. Academic studies frequently require integration of information from different texts as mentioned previously. One reason for the lack of items eliciting this process might be related to time issues.

The texts that were used in reading text V2 were long texts with 700 – 1000 words (except for search reading text which was about 3000 words). Items that require reading of the whole texts may necessitate extra time on the part of the test takers. Therefore, the timing of the test may have to be reconsidered if more items of this type were to be included.

Search reading items aimed to reveal the participants' ability in selecting information relevant to a predetermined topic. It required reading at the global and local level, using both expeditious and careful reading strategies.

The major strategies that defined search reading in the present study were reading the questions. The participants primarily read the questions, identified keywords in the questions, and returned to the questions frequently for clarification of reading aim. This finding is consistent with the literature: Weir et al. (2000) mention that in a prototype of the Advanced English Reading Test (AERT) in China, in the search reading section, the readers processed the questions before reading the text. Establishing the topic of the question, the reader then returns to the text to search for relevant information. Once they identify the relevant paragraph, they read carefully to find the answer to the question.

The pattern observed in Reading Test V2 was a similar one: high scoring participants started with the questions and used a number of strategies in an integrated manner. They used careful reading, search reading, identified the section relevant to the question using subtitles or topic, and identified the answer through sentence or paragraph meaning. Low scoring participants' pattern was slightly different: they relied more on scanning (in the sense that they tried to identify the relevant section of the text through keywords, rather than topic).

The proportion of the search reading items in the whole reading test was 27%. Across the test, the range of careful (53%) and search reading items and the use of cognitive processes revealed adequate sampling of the test construct, whereas for expeditious reading, especially expeditious global reading, there is room for a better representation of academic reading ability in the reading test.

In defining the stages of test development, the present study employed the reading model adapted from Khalifa and Weir (2009). This theory and needs driven model of reading ability for academic study guided the development of the reading test.

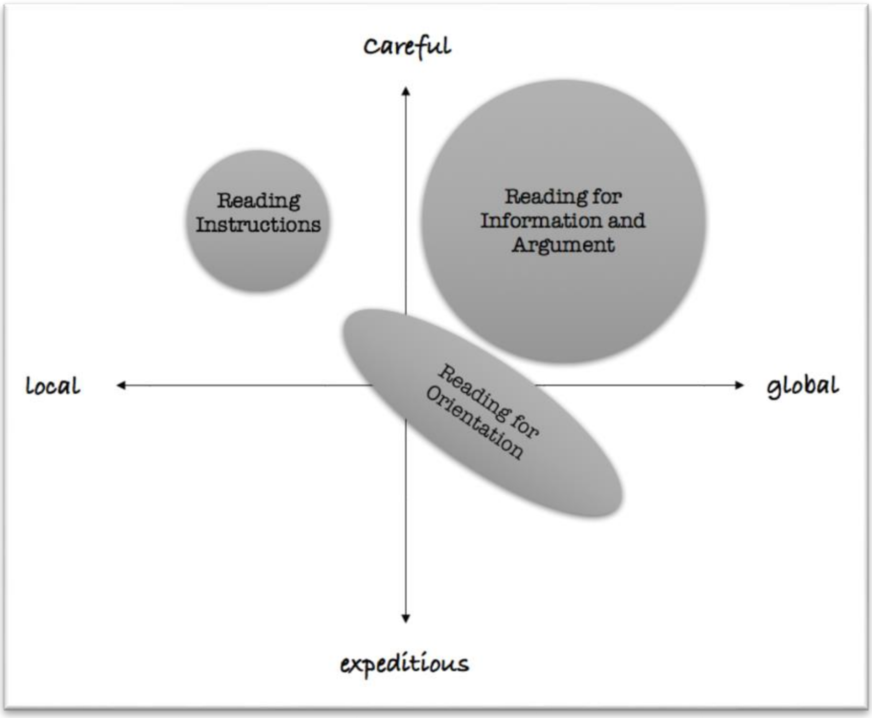


Figure 30 Theory and needs driven model

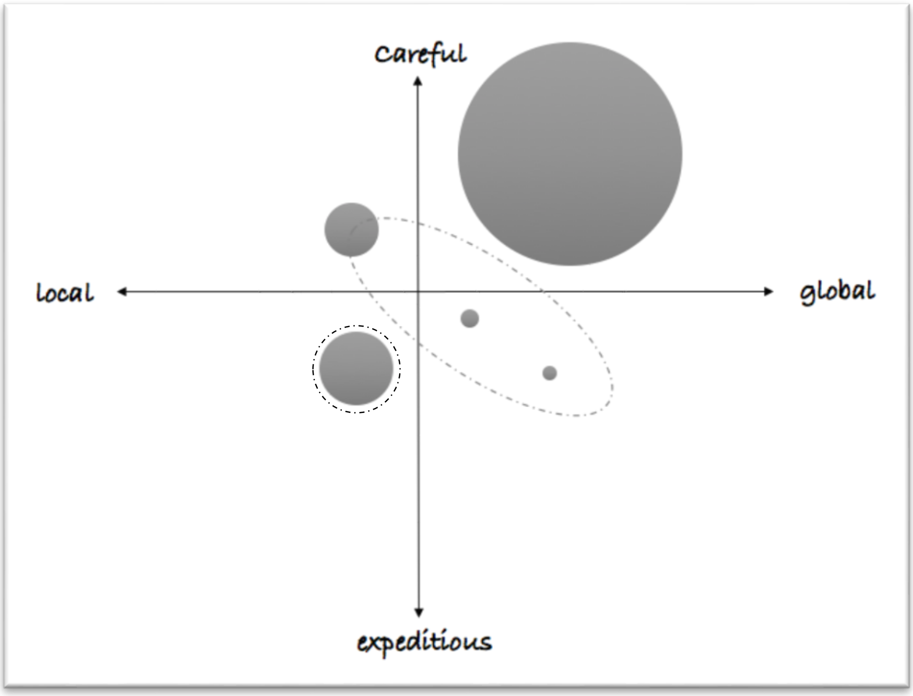


Figure 31 Data -driven reading model

After administering two different versions of the test, and evaluating the results of the study, a new model was drawn using the frequency values of the strategies used by high scoring participants in Table 51.

This data-driven model reveals that, the reading test (V2) tapped on the expected aspects of reading on two dimensions: careful / expeditious and global / local reading. As can be seen in Table 51, the primary testing objective was careful global reading, followed by local expeditious reading, and search reading (global expeditious, careful local and careful global reading). Testing of local expeditious reading was not within the suggested objectives of the test, but especially in matching and also in search reading (to a lesser extent) the participants used the scanning technique to locate specific information such as numbers or proper names. That's the reason for the appearance of the dotted circle in local expeditious reading domain. As the use of strategies during test taking take place in an integrated manner rather than discrete moves, it may not be possible to overrule the use the local expeditious strategy altogether. Therefore, this reading type will necessarily be present in the data-driven model.

Another unexpected finding is that; expeditious reading was insufficiently sampled in the test. Urquhart and Weir (1998) claim that the difference between L1 and L2 reader is most apparent in expeditious reading. To be able to claim that the reading test samples sufficiently from the test construct, it is essential that the all aspects of reading ability, as specified through theory and needs analysis study, be represented in the test. Therefore, new tasks testing expeditious reading strategies will need to be added in the next compilation of the reading test.

The data-driven model, though still unripe, is a successful representation of the reading construct in Reading Test V2. Overall, it is consistent with the theory-driven model and is amenable to improvement in the future compilations of the reading test.

5.4 Discussion of Research Question 3

The third research question aimed to reveal the extent to which item parameters contribute to the validity claims of the reading test. Bachman (1990, p. 18) defines measurement as “a process of assigning numbers to attributes of individuals or groups according to specific rules and procedures”. Within this measurement process, first the attribute is defined conceptually (theoretical basis of the attribute), then the attribute is defined operationally so that we can link the unobservable attributes to observations of performance, and finally, we quantify those observations (Bachman, 2004). This third research question is related to the properties of the numbers yielded through the quantification of those observations in the reading test. Nonetheless, these numbers are only meaningful with relation to the construct underlying the test and the operational definition of that construct. Therefore, to be able to make claims over the validity of the interpretations based on test scores, it is crucial to examine item parameters, which are a function of test scores.

Score distribution of a test is important in achieving the intended purpose of a test and the scores reveal whether the test is at the appropriate level of difficulty with relation to the ability level of the test takers. Control over the score distribution of a test is largely dependent upon control over the statistical characteristics of items (Bachman, 2004). As such, a number of statistical analyses are conducted on the test as a whole and on individual items.

The descriptive statistics on the scores of the reading test (V2) as a whole revealed that the difficulty level of the test is suitable for the intended purposes: the mean of the test is 19.8 (in other words 66%), which is only 10% higher than the administrative decision of the 60% cut-off score (that the items can be answered correctly by 60% of the test takers). This difference might be due to the fact that test takers from the pre-intermediate and pilot pre-intermediate groups were not included in the administration of the test (V2) (the reasons for this were given in Section 4.4.2).

Next to the mean score, the standard deviation (SD) was also calculated to reveal the dispersion of scores, i.e. how far each score deviates from the mean. The SD for the test was calculated to be 4.70. As the test scores were normally distributed (see Table 42), we understand that about 68% of the population scored one SD above or below the mean score, i.e. 68% of the test takers (34% + 34%) scored between 15.1 – 24.5 (in other words, between 50% - 82%) (Figure 32).

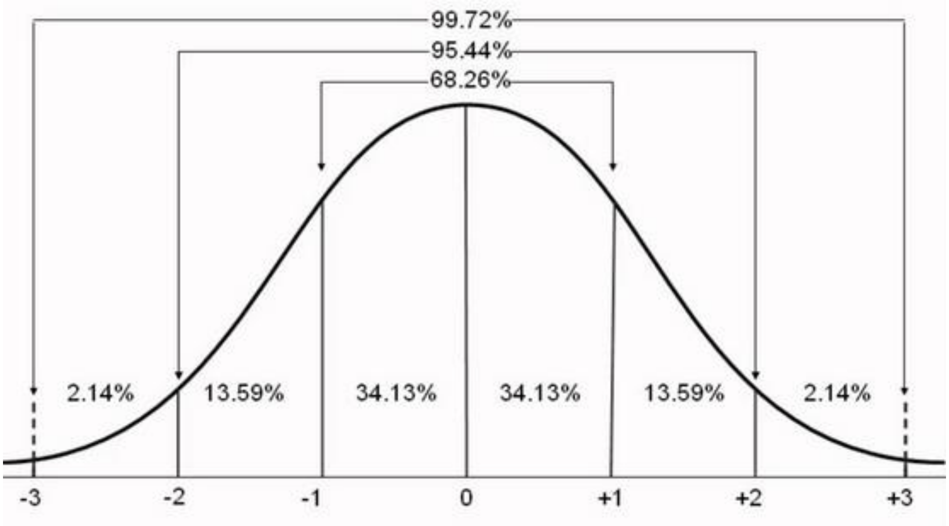


Figure 32 Normal distribution curve

The mean and SD on this test provides a reference point for us for the future administrations (or other versions) of the test. The mean score and the SD can be used to transform each score on the test (i.e. raw scores) into standardized scores, which can easily be used to compare scores with those from other administrations of the tests to reveal how different the test taker populations are from each other (J. D. Brown, 2005).

Closely related to the mean score of the test is the item facility index (IF). The IF values revealed that there were seven items that were easier, and four items that were more difficult than anticipated. The remaining 19 items were found to be within the range of 0.40 - 0.80 that had been set during test design stage.

One reason for the deviation from the 0.30 - 0.70 range that is commonly referred to in the literature (Brown, 2012; Henning, 1987) is due to the fact that the reading test is criterion-referenced in nature, as opposed to norm-referenced. In norm-referenced tests, an individual's score is compared to the others' in order to reveal the position of that individual in relation to others (for example, the Higher Education Institution Exam (Yüksek Öğrenim Kurumları Sınavı)). The higher an individual scores on a norm-referenced test, the more likely she may be offered a share of scarce resources such as a place in an academic program (Fulcher, 2010).

However, in the case of the reading test presented in this study, there are no limits to the number of people who can obtain a pass score to be admitted to their programs. The scores are interpreted with reference to a criterion: by answering a certain percentage of the questions correctly they may achieve the cut-off score of 60, which was determined to be the minimum passing score for the test by the school administration. From this perspective, the number of test takers who obtain the skills and strategies to carry out the relevant reading operations is not limited. Anyone answering a certain amount of questions correctly, the difficulty of which range between 0.40 - 0.80, achieves a passing score. The reading test is criterion-referenced as are many high-stakes tests developed by international institutions (such as TOEFL, or IELTS).

IF values closer to the mid-point of 0.50 are known to maximize variability of scores. However, according to Hambleton and Novick (1973) "a criterion-referenced test is not constructed specifically to maximize the variability of test scores (whereas a norm-referenced test is)" (p.162), which justifies the use of a higher range of IF values than is recommended in the literature (for norm-referenced tests).

Another statistic carried out using test scores was item discrimination. The discrimination indices of 19 items were within the expected range whereas 11 had values lower than the expected 0.20 limit. Almost all those items which were too easy (with an IF above 0.80) also had low discrimination indices. As more number of test takers manage to answer an item, the item's discriminative power diminishes. This is also apparent in point biserial correlation values of the same items. A study on CTT revealed that if the variance of test scores are not wide, than reliability estimates will be low (Lord & Novick, 1968). Hambleton and Novick (1973) also warn that the reliability estimates in criterion-referenced test will be low. In the light of the suggestions from the literature, and the assumptions of Cronbach's alpha coefficient estimates (that the coefficients should be calculated for items tapping on one trait/ability) calculating reliability estimates separately for each reading ability/type as was done for Reading Test V2 has been justified.

Overall, the item parameters, IF and discrimination indices, as well as the reliability estimates of the items reveal that the majority of the items reveal properties that support the decisions made based on test scores. Except for a few items that had high IF values (over 0.80, but especially those two with an IF of 0.96), the majority of the items provided ample information about the candidate's ability level. The evidence provided here is consistent with the scoring validity claims of the framework upon which the test was modeled.

5.5 Conclusion

This study validated aspects of the reading test that is part of the English Proficiency Exam at METU. It sought to generate evidence for context validity, cognitive validity and scoring validity to be able to argue that the score-based interpretations about the test takers' reading ability are justifiable. As in all high-stakes tests, score-based decisions that are not supported by sound analysis or scientific data cannot be accepted to be valid. It follows that it is the responsibility of the administering institution to provide evidence for the fairness and meaningfulness of their decisions.

Messick's seminal work on validity (1989b) provided the basis for many schemes for validity studies. His work had a significant influence on approaches to validity: rather than seeing it as a property of a test, it was defined as the extent to which we are justified in making inferences or giving decisions based on the test scores.

Messick (1989b) proposed a framework that integrated the factors related to content, criteria and consequences of a test into a comprehensive framework. This progressive framework contained four cells with facets as sources of evidence that contributed to this unified view of validity (Table 52). Using this matrix, evidence needs to be gathered for the potential and actual consequences of test score meaning and test score use.

Table 52 Messick's framework

	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

Messick's work on validity changed the focus from the test to the arguments about the test. He claims that the argument about the test justifies the inferences rather than the test itself. Therefore, "validity is a matter of degree, not all or none" (Messick, 1989b, p. 13).

Weir (2005), building upon Messick's (1989b) unitary validity concept, proposed a systematic approach to test validation through a sociocognitive framework. This framework, which can be used both for test development and validation, contains validity schemes for each stage of test development and administration. In *a priori* validation stage test-taker characteristics, contextual parameters (contextual validity) and the cognitive processes elicited by the test (cognitive validity) are examined. In the *a posteriori* stage, scoring procedures (scoring validity), social consequences of testing (consequential validity) and predictive value of the test (predictive validity) are mentioned as aspects for validation. This study focused on three aspects of the framework; namely, context validity, cognitive validity and scoring validity, which constitute the overarching concept of *construct validity* (Messick, 1995).

In examining the context validity of the reading test, the existing literature on the definitions of reading ability were reviewed. The plethora of approaches towards reading mainly fell into two categories: the process and componential views of reading. The former focus on describing the actual cognitive processes that take place during reading whereas the latter attempts to describe the subskills that are believed to underlie reading ability. The present study employed a model of reading that is based on the cognitive processing approach, following Khalifa and Weir's (2009) reading model. In this model there are three interlinked parts called **Metacognitive Activity,** which defines the type of activities that the reader carries out, **Central Processing Core,** which includes elements initiated by the activities carried out in the Metacognitive Activity, and **Knowledge Base,** which refers to the types of knowledge that the reader brings into the reading process.

The metacognitive activity contains the goal setter, which is the agent that determines the purpose for reading and decides what type of reading will be carried out to achieve that purpose. Urquhart and Weir (1998) propose two types of reading: careful and expeditious reading which are carried out either at the local or global level. In the present study, the cognitive processing model of reading was used to describe careful and expeditious reading.

After establishing the conceptual basis of reading ability, a needs analysis project was used to bring light into the specific reading requirements in the first year of undergraduate programs at the faculties at METU. This information was used to finalize the conceptual model of reading, and define the criteria to operationalize the test construct. The criterial parameters were presented in the test specifications document. They included information on task setting and linguistics demands of the test. Task setting is about the aspects of the test such as response method, time constraints and weighting of items. Linguistics demands, on the other hand, focus on the knowledge sources that the test taker uses to answer the test items such as grammatical and lexical resources. This test specifications document serves as part of the *a priori* validation of the test revealing information on context validity, and answers the first research question of this study.

In the next step, *a priori* validation, evidence was generated for the cognitive validity of the test. To this end, two versions of the test were administered: the first version was administered in parts (Subtest 1, 2, 3 and 4) to students at the DBE. About 100 participants answered each subtest. The participants marked the processes they carried out on a retrospective protocol form after answering each question in the reading test. The data from the protocol forms were analyzed quantitatively to reveal that mainly careful reading at the global level, followed by careful reading at the local level were carried out by the participants. This finding suggests that the macro comprehension and critical items were responded as the test writers intended; however, the use of expeditious reading strategies was limited, and therefore, in the next version, attention was given to developing items that elicit expeditious reading behavior.

The second version of the test was administered to 26 participants, and a think aloud protocol was carried out: each participant was asked to verbalize their thoughts while answering the questions. The major findings of the introspective protocol were that the participants used a wide variety of skills and strategies while answering the test items, which proves that the test scores can be generalized beyond the testing situation. In terms of the types of reading, similar to the findings of the first version of the reading test, careful reading was again the most frequently used reading type. With relation to the cognitive processing reading model, both lower level (e.g. decoding, understanding lexis, syntax) and higher level (e.g. inferencing) processes were carried out during the test. A newly added section to Reading Test V2, the search reading section, proved to be effective in eliciting expeditious reading strategies. Requiring a different type of response from the test-takers (i.e. short answer format), this section improved the reliability of scores as the questions tapped on a different aspect of reading (Lee, 2005).

The quantifying of the verbal protocols revealed that careful reading strategies were used more than three times as much as expeditious reading strategies. This finding is consistent with the literature: many tests developed by international organizations (such as the IELTS) emphasize careful reading at the local and global level. However, following Weir's (1983) claims, I argue that more attention should be given to the testing of expeditious reading strategies.

The cognitive processes that were employed in both versions of the test revealed findings similar to those in the literature: participants who were admitted to the DBE at the beginner or elementary level relied mainly on processes that are categorized as lower level processes (bottom-up processes), whereas those started receiving English instruction at the intermediate or upper-intermediate level at the DBE were able to activate higher level processes (such as building a mental model) more frequently and more successfully. The literature corroborates with these findings with the view that as the readers' proficiency in language increases, they carry out lower level processes in an automated manner, and concentrate better on processes that have higher cognitive load (Khalifa and Weir, 2009).

The findings of the verbal protocols provided evidence for the cognitive validity of the test and answered the second research question.

The final research question investigated the scoring validity of the test by revealing the extent to which item parameters contribute to the validity claims of the reading test. Item parameters are a function of test scores, and therefore, it is crucial to analyze the items to attribute meaning to the scores.

The descriptive statistics on the scores of the second version of the reading test revealed that despite some unfavorable item facility values, the overall difficulty level of the test was found to be suitable for the intended purposes. Another statistical property that was investigated was item discrimination indices. More than half of the items discriminated well between participants at different proficiency levels. Those items that had low discrimination indices (that do not discriminate well) were evaluated in accordance with the criterion-referenced testing norms, and some of them were marked for reevaluation with criterion-referenced testing norms (which suggest that if the variance of test scores are not wide, than reliability estimates may be low (Lord & Novick, 1968).

In brief, the item parameters revealed that the majority of the items have properties that support the decisions made based on test scores and the evidence provided here is consistent with the scoring validity claims of the framework upon which the test was modeled.

5.5.1 Reflection on validation, consequences and expectations. The present study, which covers a time span of about two years, is a one of a kind effort in test validation at METU context. Though the METU-EPE has been in effect (in slightly different forms) for at least 30 years, apart from the statistical analyses on test scores, to the best of my knowledge, there has been no validity investigation on the test as a whole or in parts since 1999 (there is a Master's thesis by Ataman (1999)). Considering the fact that it is a high stakes test administered to more than 10 thousand test takers every year, the lack of validity research on the test overshadowed its reputation, and the form, format, and content of the test have been repeatedly questioned by the stakeholders over the years. Therefore, as a unique effort to ensure the validity of various aspects of the reading section of METU-EPE, this study is a valuable contribution to the literature.

In any high-stakes testing environment, it is necessary to consider the impact of the test on the stakeholders. The more informed stakeholders the less likely there will be negative unintended consequences of test use.

The stakeholders are those who are directly or indirectly influenced by the use of the test or the interpretations of test scores. Caines et al. (2014) categorize three main stakeholders in high-stakes testing in the following manner: *test makers, test takers* and *decision makers*. In case of METU-EPE, *test makers* are the testing committee, I, as the representative of the R&D unit, DBE instructors, MLD instructors and the SFL administration. The *test takers* are DBE students and graduate candidates. The *decision makers* (and also those who rely on test results to meet their goals) are DBE and MLD instructors, faculty members, freshman students, the registrar's office, and the university board. Other stakeholders that may be considered as a fourth group are families/parents of test takers, educators external to the test-administering institution, and the society at large. Having such a large stakeholder group, the issues of validity and fairness become more urgent, as well as the issue of unintended consequences of the use of the test.

Before going back to the topic of consequences of test use, I want to repeat a piece of information I mentioned in Section 1.5: People involved in test design and development carry the burden of accountability whereas not all of them are equipped with the necessary skills and knowledge (Hatipoğlu, 2010; Taylor, 2009). In a similar vein, the *decision makers* (e.g. the registrar's office, the student affairs office, the university board) might not have the knowledge to understand what a certain score represents, "and this is exacerbated by a general lack of understanding about the imprecision or error inherent in any measurement" (Taylor, 2009, p. 22). Similarly, the public may have unrealistic ideas as to what a certain score means. It is probable that this lack of assessment literacy in general may cause issues that could be categorized as unintended consequences of test use and test interpretation.

Messick (1989b) dealt with the notion of unintended social consequences of test use in his framework (Table 52) under consequential basis of test use. The unintended consequences of test use could be either positive or negative. Some of the positive consequences mentioned in the literature are increased teacher professional development, better alignment of instruction with standards, and more remediation for low-achieving students (Cizek, 2001). On the other hand, there may be negative consequences, some of which are narrowing of the curriculum, test anxiety, pressure on the teacher (Cizek, 2011), teaching to the test, and differences in score distributions (based on, for example, gender or ethnic background) (Shepard, 1997). It is not within the scope of this study to detail the possible washback effects of the reading test on the stakeholders (although, it is an important point to consider for future studies); however, we may be cautioned about the need to increase assessment literacy among the stakeholders in order to reduce the negative effects of test use.

Popham (2011, p. 267) describes assessment literacy as consisting of "an individual's understandings of the fundamental assessment concepts and procedures deemed likely to influence educational decisions." There is consensus in the literature that the language testing community failed to encourage the public to understand and be more engaged in assessment principles and practices (Taylor, 2009). Considering my own involvement in test design/development and validation, and the complex and intricate nexus of relations between the stakeholders, it is imperative that the stakeholders increase their level of assessment literacy relative to the level of decisions they make on test scores.

5.5.2 Implications. To the best of my knowledge, the present study is the first attempt to validate aspects of the reading test of the English Proficiency Exam at METU by a) providing the conceptual and operational definitions of the test construct, b) elucidating the cognitive processes that are elicited by the test items and c) examining the item parameters to check them against established values for high-stakes tests. Acknowledging its limitations, this study has important implications for test development and validation.

5.5.2.1 Theory. To date, defining reading has been difficult. Although there seems to be a consensus on its being a cognitive activity (Martin, 1988; Stauffer, 1967; Urquhart & Weir, 1998), there are numerous approaches towards explaining the nature of reading comprehension for teaching, learning and assessment. Recently, the focus of research on reading has shifted towards reading as a process (Anderson, 1991) rather than a product.

The cognitive processing model of reading (Khalifa & Weir, 2009; Urquhart & Weir, 1998; Weir & Khalifa, 2008a) provides a methodological approach allowing the examination of the reading processes (as in careful / expeditious and local / global reading), metacognitive activities and knowledge bases that are utilized during reading (see Section 4.2.1 for details of the model). This model emphasizes not only reading in the traditional sense (that is, careful, incremental reading for comprehension), but also expeditious reading skills which were found to be of importance for academic study (Cohen & Upton, 2006; Weir, 1983). This was also verified through the local needs analysis study, and operationalization of expeditious reading was included in the test specifications document (covering all aspects of expeditious reading: skimming, scanning and search reading). Some well-known external tests such as IELTS or TOEFL have not yet included search reading in their academic reading tests. This might be due to difficulties in contextualizing the tests as they are administered in a wide variety of countries to people with different backgrounds[4]. In this respect, the present study contributes to the literature by demonstrating how to utilize a theoretical model in the design of an assessment instrument in order to address the emerging academic needs

[4] On the IELTS website – www.ielts.org – it has been stated that the test is administered in more than 140 countries, and on the ETS website – www.ets.org – it has been claimed that more than 35 million people have taken the TOEFL exam (as of 2018).

and contextual requirements. The structure and the rationale of the test has been strongly reinforced with theory and contextual needs, and as such, it provides a successful model in the field of assessment.

In addition, this study provides empirical evidence for the reading processes hypothesized in the reading model by Urquhart and Weir (1998). The purposes of reading specified in the four-cell matrix of the model were targeted in the assessment battery, and the results demonstrated that this model is meaningful in operationalizing academic reading. In this perspective, the model was found be valid to define academic reading and can be used in the design of both instructional and evaluative materials in second language contexts.

Moreover, since Urquhart and Weir's (1998) process model of reading was not used in isolation, but within a framework that helped to establish relevant parameters for the testing of reading ability in the local context, the framework provided a scheme for before and after the test events (*a priori* and *a posteriori* validation) which are argued to be the most important components in defining the test construct (Khalifa & Weir, 2009). The criterial parameters defining the test construct should reflect those that are minimally required for achievement in the target language use situation. In case of the Reading Test V2, the findings revealed that the parameters related to context and cognitive validity investigations (for example, task settings and linguistic demands for context validity, and internal processes for cognitive validity), which are derived from the literature and merged with local needs, are valid indicators of the reading operations in real-life. Therefore, this study contributes to the knowledge base by describing and illustrating the parameters of a validation framework in the development of a reading test used in academic context. This information will be useful to test developers and policy makers alike as the inferences based on test scores will have been grounded on reliable and valid criteria that are empirically justified.

Finally, the findings provide evidence for the validity of the reading model in an L2 context that was originally based on an L1 reading ability. The reading model used in this study (Khalifa and Weir, 2009) was derived from a study that explicated the reading processes taking place while reading in the native language. Applying this model to an English as a foreign language (EFL) context, this study provides evidence for the validity of the L1 reading model for L2.

5.5.2.2 Assessment practice. The results of METU-EPE, as well as some other assessment instruments such as TOEFL, are accepted at METU as proof of a certain level of English language proficiency. The registrar's office relies on these results to make decisions about candidates such as granting admission to an academic program. These results are expected to reflect the candidate's communicative abilities that are needed to achieve success in academic programs. If the scores are not truly reflective of the required abilities then both the candidate and the institution will suffer the consequences. To name a few, admitting unqualified candidates to a program may disrupt the teaching and learning environment, the candidates' failure as a result of the use of inappropriate measurement instruments may cost them time, money, and other resources, and the failing cohort may reflect badly on the institution.

The key to making informed and accurate score-based decisions is to ensure that the assessment instrument measures abilities relevant to the purpose of the test and at an appropriate level, i.e. contains test tasks that replicate communicative activities of the target language domain, and the level is set appropriately. To do so, the test construct should be defined carefully and thoroughly.

This study set out to define the test construct, and to set criterial parameters relevant in the target language use domain through a framework. The test construct adequately samples from the reading ability and the criteria for success were set in accordance with the local needs and suggestions in the literature. The definition of the reading test construct was based on Urquhart and Weir's (1998) matrix of reader purpose, and both careful and expeditious reading at the local and global level were aimed with the test items. The test scores, therefore, correctly reflect test takers' ability on aspects of reading that are important in the academic context. Following Messick's (1989b) definition of validity, the empirical evidence and theoretical rationale of the reading test support the adequacy and interpretations made on test scores. This test development model, therefore, provides a systematic and sound approach to test development, and can be used in similar contexts.

5.5.2.3 Instructional practice. The reading test developed through the stages specified in the validation framework was administered to students of the language school at METU. Therefore, the whole test development process provides important information to instructors who teach reading skills to students at the language school. First of all, the scores from the test reveal to what extent reading instruction at the school is successful, with relation to the expectations at the academic programs at METU as well as which aspects of the instructional program with relation to the teaching of reading fail or do not fully serve their purpose. The results of the test can be used to improve the relevant aspects of the curriculum such as reading objectives and learning outcomes. Reading syllabi can be designed based on the same reading model used in this study: the cognitive processing model of reading. This model serves as the theory of reading underlying the design of materials, tasks, and instructional activities.

Secondly, through the test scores students' needs will be made more explicit, and based on those needs, specialized instruction can be offered. Students at different ability levels can receive instruction that targets needs relevant to their level. This study has shown that learners at the lower proficiency levels rely mainly on bottom-up processes in reading whereas those learners with higher proficiency are more successful in using the interactive approach; that is, they carry out both bottom-up and top-down processes simultaneously to make meaning of the text. Materials developers are advised to design the materials using criteria specified in this study to set the difficulty level appropriately for the learners.

Thirdly, the design of the reading test in this meticulous manner, considering various aspects of the context, the features of the test taker and the cognitive processes involved in the process of reading, offers a systematic method for the teaching of reading skill as well. Instruction that is grounded in theory and supported with established needs can help students be equipped with skills and knowledge required for successful academic reading. The reading model reveals the reading types necessary for academic reading: expeditious and careful reading. The instructors are therefore, advised to provide enough opportunities to the learners to learn and practice both types of reading for academic study.

Finally, the criterial parameters set for the reading construct can be used as performance indicators by the instructors. This may help in the instruction, as well as the design of materials, and assessment instruments. For example, reading speed, length of texts, and the range of vocabulary needed for first year academic studies are specified in the test specifications document. Grounding materials and instruction on relevant criteria will provide realistic outcomes, which can be better assessed and feedback can become more relevant and better targeted.

Washback is used to refer to the positive or negative impact of testing on the stakeholders (i.e., learners, teachers, and administrators) as well as the process of learning and teaching. Messick (1989b) places this notion of the impact of tests and testing within his unified concept of validity, and names it social consequences of testing (see Section 2.2.5). According to his theory, tasks that are authentic and that replicate real-life activities promote positive washback. Therefore, the contribution of this study to the teaching and testing of reading is clear-cut: it makes explicit methods of test design and development, and provides information on all facets of the test construct, which can be used both in the testing and teaching of reading.

5.5.2.4 Administrative approach. Institutions that develop their own tests have the responsibility provide evidence that the interpretations they make based on test scores are justifiable. In order to do this, a systematic approach in the design and development of tests is needed. This study demonstrates a viable and reliable approach in test design and development through the use of a framework: the temporal sequencing of the framework provides information on the types of investigation needed for each stage of test development. Policy makers and administrators at schools are recommended to establish a similar approach in instruction as well as in assessment decisions.

For the teaching of reading, as well as that of other skills, this study provides a systematic method to administrators at language schools within universities: the skill should be defined taking into consideration the context, the learners and the theoretical implications. In this study, reading was defined based on a cognitive processing model, which proved to be successful in the development of the reading test. The administrators are advised to implement a similar approach both in program development and the design of assessment instruments. For the latter, the purpose of the test should be established clearly, and then the abilities to be tested should be defined (conceptual and operational definitions). This study provides criterial parameters that could be used in defining the observable behaviour for the reading skill.

5.5.3 Limitations of the study.

5.5.3.1 Sampling. The first limitation of the study is about the sampling method and sample numbers. The data for the study was collected in two stages: in stage one, 400 students from the DBE were involved. Each subtest of Reading Test V1 was administered to around 100 students with an even distribution from the five level groups at the DBE (the pre-intermediate, intermediate, upper-intermediate, advanced and repeat groups). The sampling used was random stratified sampling: for each subtest, one class from each level group was randomly chosen (a total of 20 classes), and as such, the number of participants and the sampling were appropriate. However, in stage two, due to the data collection method (verbal reports) a much smaller sampling could be done and participants were recruited; therefore, it was a biased sample.

The verbal report from each participant comprised of about two-hour recording. At the end of data collection, there was about 55 hours of recording for transcription and coding. As the data analysis was carried out by me alone, this amount of data was as much as could be reliable handled.

Studies investigating cognitive processes report smaller sample sizes. For example, Buck (1991) investigated cognitive processes during a listening test with six participants, and Shin (2006) investigated the construct validity of listening test items through verbal protocols using eight participants. In comparison, in the present study, 26 participants' verbal reports were analyzed. Still, the data from these participants may be not be considered sufficient for the generalizability of the results.

Another limitation concerning sampling is about the fact that the pre-intermediate level group students were not included in the second phase of the study, with the assumption that those students were not yet at the required level of proficiency to take the test. Those students were going to complete the instructional period at the end of July whereas the data was collected in April-May. Since they had not yet received the knowledge and skills to sit a reading test at the assumed B2 level, and had not yet received training in test-taking in the proficiency test format, the data could have been muddied. Waiting till the end of their instructional period was not possible due to the fact that all the other participants would have left school by that time. For this reason, the students at the pre-intermediate and repeat groups were left outside the study. However, for a test designed as part of a high-stakes test, it is especially important to understand how the borderline level candidates would perform. Therefore, it is crucial to carry out further research to examine the test results and cognitive processes of all level groups for the evaluation of the reading test.

Yet another limitation regarding the participants is about the homogenous nature of participants' backgrounds. In stage one, despite the fact that the participants were chosen by stratified random sampling, the majority were Turkish. Considering that the ratio of international students at METU is about 5-10% depending on the faculty, the sample did not successfully represent the population. This may partly be due to the fact that the majority of the international students have already been equipped with English language skills and therefore they did not study at the DBE, which was the universe for the sampling. In stage two, the participants were recruited through flyers written in English and posted at the entrance doors to the school buildings. No international student applied to take part in the study. Therefore, all the participants in stage two were Turkish. In testing, it is important to avoid bias, be it gender, or cultural bias. By having a homogenous participants group, I was not able to check whether the test was biased against a certain group of people. Before the implementation of the reading test, it is necessary to check for bias by including a higher number of participants from different backgrounds.

5.5.3.2 Data collection method. Two types of data were collected in the present study: retrospective and introspective verbal protocols, both of which were self-reports. Using self-reported data has obvious advantages for the purposes of the present study, which sought to reveal the types of cognitive processes that were activated during the answering of a reading test. However, it is also known that in self-report there is reliance on the honesty of the participants, and on the ability of the participants to accurately verbalize their thought processes. Since the present study involved the investigation of abstract mechanisms, the participants' personal and subjective understandings and perceptions might have had an effect on how they verbalized the processes. Therefore, in the interpretation of the findings, this point should be considered.

Moreover, self-reported data produces ordinal data, which can only be rank ordered but cannot be subjected to many statistical analyses. In essence, self-reported data as used in the present study is more about defining abstract concepts rather than revealing similarities or differences between the concepts. For detailed statistical analysis, another administration of the test to a large sample will be needed.

Another limitation regarding data collection is about the think aloud procedure. As I have mentioned previously (see Section 3.3.2.3.1.3), before I started recording participants' thoughts during the test, I demonstrated how to do the procedure using a sample question: I read the question, and while searching for the answer, I verbalized my thoughts. My demonstration, that is, the way I think about a question and the way I search for an answer is, obviously, one of the many possible approaches. However, it is probable that the participants may have been influenced by my 'style' of thinking aloud, and may have adopted a similar approach rather than verbalizing their own personal style in approaching and answering a question.

5.5.3.3 Data analysis methods.

5.5.3.3.1 Qualitative. The verbal data from the administration of Reading Test V2 were recorded on a tape, and then transcribed and coded using a rubric from the literature. The coding was carried out by me alone. In order to increase the reliability of coding, I recoded five of the transcriptions some time after the initial coding, and compared the results with the first coding. There was 5-10% discrepancy between the two codings. Although this ratio is low, a more favorable manner to conduct the analysis of this type of data is to follow Lincoln and Guba (1985), who propose the inclusion of a second researcher. They suggest that in qualitative research, achieving confirmability (the degree to which the results are confirmed by others) is possible with the help of another researcher. As I did not have this chance during the course of this study, I can recommend further studies to try to achieve confirmability by including another researcher to confirm that the analysis was carried out appropriately.

5.5.3.3.2 Quantitative. The answers to the questions in Reading Test V1 and V2 were analyzed using CTT conventions, i.e. item facility values, item discrimination indices and internal reliability values were examined. Although CTT analyses are used extensively in language testing, their dependence on the population is also well-known. The item facility values are only true for the population whose scores are analyzed. A better and more reliable method for examining item parameters is the Item Response Theory (IRT) which looks at the relation between the ability that is being tested and the difficulty of the test. It is a more sophisticated and flexible approach than the CTT (Thissen & Steinberg, 1988). However, IRT requires larger sample numbers (minimum sample size is 250 for high stakes tests) which made it impossible to use in the analysis of participant responses (Linacre, 1994). It is, however, advisable to use IRT after the actual administration of the test to examine its difficulty level independent of the test taking population.

5.5.4 Future research. This study set out to examine aspects of validity in a reading test through the use of a framework and a theoretical ability model within a specific context. In order to examine the validity of the framework, a similar study needs to be carried out in other contexts with similar purposes. It will reveal if the framework used in this study can appropriately accommodate contextual parameters different from those in the current study. Such a study would further demonstrate the validity of the framework for use in test development and validation.

A second line of research that I recommend is on the alignment of the reading test with an international standard. There are a number of ways to do this. One approach could be to give this test and a well-established external test within a short time period to the same cohort. The comparison of the results would reveal to what extent the reading test measures the intended ability reliably. Favorable results could improve the validity claims of the test.

Another international standard could be the Common European Framework of Reference (CEFR). The CEFR has already become or is becoming an industry standard in many contexts (for example in school admissions, university course requirements, and employment); this is apparent in the plethora of studies carried out with relation to language teaching or testing (see, for example, Chan, Inoue, & Taylor, 2015; Harsch & Hartig, 2015; Ilc & Stopar, 2014; Jones & Saville, 2009; Lowie et al., 2010; Martyniuk, 2010; Takala, 2010). Therefore, a research study on the alignment of the reading test with the CEFR scales would be valuable in terms of the validity claims of the test. Moreover, the results will reflect international standards and therefore, be more easily evaluated against relevant criteria.

A final but equally important future research recommendation is on the validation of the other sections of METU-EPE. As reading was considered the most important academic skill by the stakeholders (see Section 4.2.2, Freshman students' needs) the present study focused on the reading section of the METU-EPE. However, to make claims about the meaningfulness and fairness of the score interpretations of the whole test, similar validation studies should be conducted on the other sections of the test.

REFERENCES

Alagumalai, S., & Curtis, D. (2005). Classical test theory. In S. Alagumalai, D. Curtis, & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 1–14). Dordreht: Springer Netherlands.

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

Alderson, J. C. (2000b). Technology in testing: The present and the future. *System*, *28*(4), 593–603. https://doi.org/http://dx.doi.org/10.1016/S0346-251X(00)00040-3

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

Anderson, N. J. (1991). Individual differences in strategy use in second language reading and testing. *The Modern Language Journal*, *75*(4), 460–472.

Ataman, F. (1999). *A study on reliability and validity studies of METU School of Foreign Languages, Department of Basic English, June 1998 and September 1998 proficiency tests*. Middle East Technical University.

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

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*(1), 1–42. https://doi.org/10.1191/026553200675041464

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

Bachman, L. F. (2005). Building and Supporting a Case for Test Use. *Language Assessment Quarterly: An International Journal*. https://doi.org/10.1207/s15434311laq0201_1

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests. Oxford Applied Linguistics.* https://doi.org/10.2307/328718

Bachman, L. F., & Palmer, A. S. (2013). *Language assessment in practice*. Oxford: Oxford University Press.

Bannur, F. M., Abidin, S. A. Z., & Jamil, A. (2015). A validation process of ESP testing using Weir's socio cognitive framework. *Procedia - Social and Behavioral Sciences*, *202*, 199–208. https://doi.org/http://dx.doi.org/10.1016/j.sbspro.2015.08.223

Biber, D., & Gray, B. (2014). Discourse characteristics od writing and speaking task types on the TOFL IBT test: A lexico-grammatical analysis. *ETS Research Report Series*, *2013*(1), i-128. https://doi.org/10.1002/j.2333-8504.2013.tb02311.x

Brindley, G. (2001). Outcomes-based assessment in practice: Some examples and emerging insights. *Language Testing*, *18*(4), 393–407. https://doi.org/10.1177/026553220101800405

Brown, J. D. (2002). The Cronbach's alpha reliability estimate. *JALT Testing and Evaluation*, *6*(1), 17–19.

Brown, J. D. (2005). *Testing in Language Programs: A comprehensive guide to English language assessment*. *Mc Graw-Hill* (2nd Ed.). New York: Mcraw-Hill Publications.

Brown, J. D. (2012). Classical test theory. In G. Fulcher & F. Davidson (Eds.), *Routledge handbook of language testing*. Oxon and New York: Routledge.

Brown, R. H., & Goodman, D. (2001). Jurgen Habermas' Theory of Communicative Action: an incomplete project. In *Handbook of scoial theory* (pp. 201–216). https://doi.org/10.4135/978-1-84860-835-1.n16

Browne, C., Culligan, B., & Phillips, J. (2013a). The New Academic Word List. Retrieved September 16, 2015, from http://www.newacademicwordlist.org

Browne, C., Culligan, B., & Phillips, J. (2013b). The New General Service List. Retrieved September 16, 2015, from http://www.newgeneralservicelist.org

Bruce, I. (2008). *Academic writing and genre: A systematic analysis*. London: Continuum.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley, California: University of California Press.

Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing*, *8*(1), 67–91. https://doi.org/10.1177/026553229100800105

Caines, J., Bridglall, B. L., & Chatterji, M. (2014). Understanding validity and fairness issues in high-stakes individual testing situations. *Quality Assurance in Education*, *22*(1), 5–18. https://doi.org/10.1108/QAE-12-2013-0054

Carrell, P. L. (1987). Content and formal schemata in ESL reading. *Tesol Quarterly*, *21*(3), 461–481. https://doi.org/10.2307/3586498

Carrell, P. L. (1988). Introduction: Interactive approach to second language reading. In P. L. Carrell, J. Devine, & D. Eskey (Eds.), *Interactive Approaches to Second Language Reading*. New York: Cambridge University Press.

Cartwright, K. B. (2009). The role of cognitive flexibility in reading comprehension. In S. E. Israel & G. G. Duffy (Eds.), *Handbook of Research on Reading Comprehension*. New York: Routledge.

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing*, *14*(1), 3–22. https://doi.org/10.1177/026553229701400102

Chalhoub-Deville, M., Alcaya, C., & Mccollum Lozier, V. (2013). *An operational framework for constructing a computer- adaptive test of L2 reading ability: Theoretical and practical issues* (No. 612.626.8600). Minneapolis. Retrieved from www.carla.umn.edu/resources/working-papers/

Chan, S., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing*, *26*, 20–37. https://doi.org/http://dx.doi.org/10.1016/j.asw.2015.07.004

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces Between Second Language Acquisition and Language Testing Research* (pp. 32–70). Cambridge: Cambridge University Press.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, 254–272. https://doi.org/10.1017/S0267190599190135

Chapelle, C. A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing*, *20*(4), 409–439. https://doi.org/10.1191/0265532203lt266oa

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, *20*(4), 19–27. https://doi.org/10.1111/j.1745-3992.2001.tb00072.x

Cohen, A. D. (1998). Strategies in learning and using a second language.

Cohen, A. D. (2006). The coming of age of research on test-taking strategies. *Language Assessment Quarterly*, *3*(4), 307–331. https://doi.org/10.1080/15434300701333129

Cohen, A. D., & Upton, T. A. (2006). Strategies in responding to the new TOEFL reading tasks. ETS Research Report Series, 2006: i-162. doi:10.1002/j.2333-8504.2006.tb02012.x

Cohen, A. D., & Upton, T. A. (2007). “I want to go back to the text”: Response strategies on the reading subtest of the new TOEFL®. *Language Testing*, *24*(2), 209–249. https://doi.org/10.1177/0265532207076364

Cohen, L., Manion, L., & Morrison, K. (2007). *Research Methods in Education*. *Education* (Vol. 55). New York: Lawrence Erlbaum Associates Publishers.

Cohen, R. J., & Swerdlik, M. (2009). *Psychological testing and assessment: An introduction to tests and measurement* (7th Ed.). Boston: Mcraw-Hill Publications.

Corkill, A. J., Bruning, R. H., & Glover, J. A. (1988). Advance organizers: Concrete versus abstract. *Journal of Educational Research*, *82*(2), 76–81. https://doi.org/10.1080/00220671.1988.10885871

Cortina, J. M. (1993). What Is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104.

Council of Europe. (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)*. Strasbourg.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In *Test Validity* (pp. 3–17). https://doi.org/10.1017/CBO9781107415324.004

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. https://doi.org/10.1037/h0040957

Davidson, F. (2012). Test specifications and criterion referenced test development. In G. Fulcher & F. Davidson (Eds.), *Routledge handbook of language testing* (pp. 197–207). Oxon: Routledge.

Davidson, F., & Fulcher, G. (2007). The Common European Framework of Reference (CEFR) and the design of language tests : A matter of effect. *Language Teaching*, *40*, 231–241.

Davies, A., & Elder, C. (2005). Validity and validation in language testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 795–813). New Jersey: Lawrence Erlbaum Associates Publishers.

Devi, S. (2010). *Validating aspects of a model of academic reading*. University of Bedfordshire.

Donaghue, H., & Thompson, J. (2012). A reading model for foundation year students at a tertiary institution in the United Arab Emirates. *Cambridge ESOL: Research Notes*, *49*, 40–46. Retrieved from http://www.cambridgeenglish.org/images/23166-research-notes-49.pdf

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

Ellis, R. (1994). The study of second language acquisition. *Oxford Applied Linguistics*, *23*, 824.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper*. Princeton, N.J.

Farhady, H. (2012). Principles of language assessment. In C. Coombe, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second langauge assessment* (pp. 37–46). New York: Cambridge University Press.

Field, J. (1999). Key concepts in ELT. *ELT Journal*, *53*(4), 338–339. Retrieved from https://doi.org/10.1093/eltj/53.4.338

Field, J. (2004). *Psycholinguistics: The key concepts*. London: Routledge.

Fox, J. (2004). Test decisions over time: tracking validity. *Language Testing*, *21*(4), 437–465. https://doi.org/10.1191/0265532204lt292oa

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

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

Furr, M. R., & Bacharach, V. R. (2008). *Psychometrics: An introduction*. Thousand Oaks: Sage Publications.

Gamboa-González, Á. M. (2017). Reading comprehension in an English as a foreign language setting: Teaching strategies for sixth graders based on the interactive model of reading. *Folios*. scieloco.

Garrett, H. E. (1937). *Statistics in psychology and education* (2nd Ed.). Oxford: Longmans.

Grabe, W. (1991). Current Developments in Second Language Reading Research. *TESOL Quarterly*, *25*(3), 375–406. https://doi.org/10.2307/3586977

Grabe, W. (2009). *Reading in a second language Moving from theory to practice*. https://doi.org/10.1111/j.1540-4781.2011.01151.x

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspective*. London: Longman.

Grabe, W., & Stoller, F. L. (2011). *Teaching and researching reading*. Harlow: Longman/Pearson.

Green, A. (1998). Verbal protocol analysis in language testing research: A handbook. *Language Testing*, *16*, 483–486.

Green, A. (2014). *Exploring language assessment and testing: Language in action*. New York, NY: Routledge.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a Conceptual Framework for Mixed-Method Evaluation Designs. *Educational Evaluation and Policy Analysis*, *11*(3), 255–274. https://doi.org/10.2307/1163620

Grotjahn, R. (1986). Test validation and cognitive psychology: some methodological considerations. *Language Testing*, *3*(2), 159–185. https://doi.org/10.1177/026553228600300205

Grotjahn, R. (2001). TestDaF: Theoretical basis and empirical research. In M. Milanovich & C. J. Weir (Eds.), *Studies in language testing: European language testing in a global contex* (pp. 189–203). Cambridge: Cambridge University Press.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, *6*(4), 427–438. https://doi.org/10.1177/001316444600600401

Gürsoy, S. (2013). *The English proficiency exam in EFL context: A validation study*. Çağ University.

Haertel, E. (1985). Construct validity and criterion-referenced testing. *Review of Educational Research*, *55*(1), 23–46. https://doi.org/10.2307/1170406

Haladyna, T. M., & Rodriguez, M. C. (2004). *Developing and validating multiple-choice test items* (3rd Ed.). New Jersey: Lawrence Erlbaum Associates, Inc.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, *10*(3), 159–170.

Harsch, C., & Hartig, J. (2015). What are we aligning tests to when we report test alignment to the CEFR? *Language Assessment Quarterly*, *12*(4), 333–362. https://doi.org/10.1080/15434303.2015.1092545

Hatipoğlu, Ç. (2010). Summative evaluation of an English language testing and evaluation course for future English language teachers in Turkey. *ELTED*, *13*, 40–51.

Hatipoğlu, Ç. (2015). English language testing and evaluation (ELTE) training in Turkey: Expectations and needs of pre-service English language teachers. *ELT Research Journal*, *4*(2), 111–128.

Hatipoğlu, Ç. (2016). The impact of the University Entrance Exam on EFL education in Turkey: Pre-service English language teachers' perspective. *Procedia-Social and Behavioral Sciences*, *232*, 136–144.

Hegarty, M., & Just, M. A. (1989). Understanding machines from text and diagrams. *Advances in Psychology*, *58*(C), 171–194. https://doi.org/10.1016/S0166-4115(08)62154-8

Henning, G. (1987). *A guide to language testing*. Cambridge, MA: Newbury House.

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, *8*(2), 689–696. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl82hirsh.pdf

Holliday, W. G., Brunner, L. L., & Donais, E. L. (2018). Differential cognitive and affective responses to flow diagrams in science. *Journal of Research in Science Teaching*, *14*(2), 129–138. https://doi.org/10.1002/tea.3660140205

Hudson, T. (1998). Theoretical perspectives on reading. *Annual Review of Applied Linguistics*. https://doi.org/10.1017/S0267190500003470

Hyland, K. (2002). *Teaching and Researching Writing*. *Teaching and Researching Writing*. London: Longman. https://doi.org/10.4324/9781315833729

Hyland, K. (2008). Genre and academic writing in the disciplines. *Language Teaching*, *41*(4). https://doi.org/10.1017/S0261444808005235

Hyon, S. (1998). Text, role, and context: Developing academic literacies. *English for Specific Purposes*. https://doi.org/10.1016/S0889-4906(97)84493-6

Ilc, G., & Stopar, A. (2014). Validating the Slovenian national alignment to CEFR: The case of the B2 reading comprehension examination in English. *Language Testing*, *32*(4), 443–462. https://doi.org/10.1177/0265532214562098

Introduction to the CEFR with checklists of descriptors – Eaquals. (n.d.). Retrieved October 5, 2016, from https://www.eaquals.org/resources/introduction-to-the-cefr-with-checklists-of-descriptors/

Jakobson, R. (1960). Closing Statement: Linguistics and Poetics. In T. A. Sebeok (Ed.), *Style in language* (pp. 350–377). Cambridge, MA, MA: MIT Press.

Jamieson, J. (2013). Defining constructs and assessment design. In *The Companion to Language Assessment*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla062

Jenkins, J. G. (1946). Validity for what? *Journal of Consulting Psychology*, *10*(2), 93–98. https://doi.org/10.1037/h0059212

Jones, N. (2001). Reliability as one aspect of test quality. *Research Notes*. Retrieved from http://www.cambridgeenglish.org/images/23115-research-notes-04.pdf

Jones, N., & Saville, N. (2009). European language policy: Assessment, learning and the CEFR. *Annual Review of Applied Linguistics*, *29*, 51–63. https://doi.org/doi:10.1017/S0267190509090059

Jun Zhang, L. (2001). Awareness in reading: EFL students' metacognitive knowledge of reading strategies in an acquisition-poor environment. *Language Awareness*. https://doi.org/10.1080/09658410108667039

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354. https://doi.org/10.1037/0033-295X.87.4.329

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2011). Validating score interpretations and uses. *Language Testing*, *29*(1), 3–17. https://doi.org/10.1177/0265532211417210

Kane, M. T. (2012). All validity Is construct validity. Or Is It? *Measurement*, *10*(1), 66–70. https://doi.org/10.1080/15366367.2012.681977

Kane, M. T. (2013). The argument-based approach to validation. *Social Psychology Review*, *42*(4), 448–457.

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy and Practice*, *23*(2), 198–211. https://doi.org/10.1080/0969594X.2015.1060192

Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(1), 5–17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x

Kaplan, R. M., & Saccuzzo, D. P. (1982). *Psychological Testing: Principles, Applications, and Issues*. *Brooks/Cole Publishing Company*.

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*(6). https://doi.org/10.4304/tpls.3.6.877-884

Kelley, T. L. (1927). *Interpretation of educational measurements*. *Journal of Applied Psychology* (Vol. 12). https://doi.org/10.1037/h0068663

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading* (Studies in). Cambridge: UCLES/Cambridge University Press.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*. https://doi.org/10.1037/0033-295X.85.5.363

Kobayashi, M. (2002). Method effects on reading comprehension test performance: text organization and response format. *Language Testing*, *19*(2), 193–220. https://doi.org/10.1191/0265532202lt227oa

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

Kucan, L., & Beck, I. L. (1997). Thinking aloud and reading comprehension research: Inquiry, instruction, and social interaction. *Review of Educational Research*, *67*(3), 271–299. https://doi.org/10.2307/1170566

Kutevu, E. (2001). *Investigating the validity of achievement and proficiency tests at Bilkent University School of English Language*. Middle East Technical Unievrsity.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In H. Bejoint & H. Arnaud (Eds.), *Vocabulary and applied linguistics* (pp. 126–132). London: Macmillan.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge: Cambridge University Press.

Lee, Y.-W. (2005). *Dependability of scores for a new ESL speaking test: Evaluating prototype tasks*. Princeton, N.J.

Linacre, J. M. (1994). Sample size and item calibration [or person measure] stability. Retrieved from https://www.rasch.org/rmt/rmt74m.htm

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic Inquiry*. Newbury Park: Sage Publications Inc.

Lissitz, R. W. (2009). *The concept of validity: Revisions, new directions and applications*. *The Concept of Validity Revisions New Directions and Applications* (Vol. 48). https://doi.org/10.1111/j.1745-3984.2011.00155.x

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Mass: Addison-Wesley.

Lowie, W. M., Haines, K. B. J., & Jansma, P. N. (2010). Telling ELT Tales out of School Embedding the CEFR in the academic domain: Assessment of language tasks. *Procedia Social and Behavioral Sciences*, *3*, 152–161.

Lynch, B. K. (2003). *Language assessment and program evaluation*. Edinburgh: Edinburgh University Press.

Martin, S. H. (1988). A description of cognitive processes during reading and writing. *Reading Psychology*, *9*(1), 1–15. https://doi.org/10.1080/0270271880090102

Martyniuk, W. (2010). *Aligning tests with the CEFR : Reflections on using the Council of Europe's draft manual*. Cambridge University Press.

Materials for the Guidance of Test Item Writers. (n.d.). Retrieved from http://www.alte.org/attachments/files/item_writer_guidelines.pdf

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375–407. https://doi.org/10.1037/0033-295X.88.5.375

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, *3*(1), 31–51. https://doi.org/10.1207/s15434311laq0301_3

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. https://doi.org/10.3102/0013189X018002005

Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13–103). New York: Macmillan.

Messick, S. (1990). *Validity of test interpretation and use*. Princeton, NJ.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*. https://doi.org/10.1037/0003-066X.50.9.741

METU, Middle East Technical University. (n.d.). Retrieved May 13, 2014, from http://www.metu.edu.tr/

Meyer, B. J., & Freedle, R. O. (1984). Effects of discourse type on recall. *American Educational Research Journal*, *21*(1), 121–143. https://doi.org/10.3102/00028312021001121

Milanovich, M., Saville, N., Pollitt, A., & Cook, A. (1996). Developing rating scales for CASE: Theoretical concerns and analysis. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 15–38). Avon: Cromwell Press.

Mislevy, R. J. (2007). Validity by Design. *Educational Researcher*, *36*(8), 463–469. https://doi.org/10.3102/0013189X07311660

Moore, T., Morton, J., & Price, S. (2007). *Construct validity in the IELTS Academic Reading test: A comparison of reading requirements in IELTS test items and in university study*.

Moss, P. A. (2007). Reconstructing Validity. *Educational Researcher*, *36*(8), 470–476. https://doi.org/10.3102/0013189X07311608

Nakatsuhara, F. (2011). The relationship between test-takers' listening proficiency and their performance on the IELTS Speaking Test. In J. Osborne (Ed.), *IELTS Research Reports Volume 12* (pp. 1–50). Melbourne: IDP: IELTS Australia and British Council.

Nation, I. S. P. (1990). *Teaching and learning vocabulary. Handbook of research in second language teaching and learning*.

Nation, P. (2009). Reading faster. *International Journal of English Studies*, *9*(2), 131–144.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. *Vocabulary: Description, Acquisition and Pedagogy*, *14*, 6–19.

Nevo, B. (1980). Item analysis with small samples. *Applied Psychological Measurement*. https://doi.org/10.1177/014662168000400304

Nuttall, C. (1996). *Teaching reading skills in a foreign language*. London: Heinemann.

O'Sullivan, B. (2000). *Towards a model of performance in oral language testing*. University of Reading.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, *19*(3), 277–295. https://doi.org/10.1191/0265532202lt205oa

O'Sullivan, B., & Weir, C. J. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language Testing: Theories and Practices*. London: Palgrave Macmillan.

Oakhill, J., & Garnham, A. (1988). *Becoming a skilled reader*. Oxford: Blackwell Publishers Inc.

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats*. Boston: Kluwer Academic Publishers.

Pearson, D., & Kamil, M. (1978). *Basic processes and instructional practices in teaching reading*.

Pearson ELT. (n.d.). Global Scale of English.

Peirce, B. N. (1992). Demystifying the TOEFL® Reading Test. *TESOL Quarterly*, *26*(4), 665–691. https://doi.org/10.2307/3586868

Popham, W. J. (1990). *Modern educational measurement* (2nd Ed.). Englewood Cliffs. NJ: Prentice Hall.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice*, *48*(1), 4–11. https://doi.org/10.1080/00405840802577536

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator*, *46*(4), 265–273. https://doi.org/10.1080/08878730.2011.605048

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading:* New Jersey: Lawrence Erlbaum Associates, Inc.

Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling*, *52*(4), 354–379.

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, *18*(1), 1–32. https://doi.org/10.1177/026553220101800101

Reynolds, C. R., & Suzuki, L. A. (2003). Bias in psychological assessment. In J. R. (John R. Graham, J. A. Naglieri, & I. B. Weiner (Eds.), *Handbook of psychology. Volume 10, Assessment psychology* (pp. 67–89). New Jersey: Wiley.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13.

Rosenfeld, M., Leung, S., & Oltman, P. K. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels*. New Jersey.

Rost, M. (2013). *Teaching and researching listening, second edition*. *Teaching and Researching Listening, Second Edition*. https://doi.org/10.4324/9781315833705

Rumelhart, D. E. (1977). Toward an interactive model of reading. In S. Dornic (Ed.), *Attention and performance VI: proceedings of the Sixth International Symposium on Attention and Performance* (pp. 573–606). New Jersey: Lawrence Erlbaum Associates Publishers.

Rupp, A. a, Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: a cognitive processing perspective. *Language Testing*. https://doi.org/10.1191/0265532206lt337oa

Saville, N. (2003). The process of test development and revision within UCLES EFL. In C. J. Weir & M. Milanovich (Eds.), *Continuity and innovation: Revising the Cambridge Proficiency in English Examination 1913 - 2002* (pp. 57–120). Cambridge: UCLES/Cambridge University Press.

Schmitt, N., & McCarthy, M. (1997). *Vocabulary: Description, acquisition and pedagogy*. Cambridge University Press.

SFL. (2015). *Program evaluation and needs analysis project*. Unpublished manuscript, SFL, METU, Ankara, Turkey.

Shaw, S. D., & Weir, C. J. (2008). *Studies in language testing: Examining writing* (2nd Ed.). Cambridge: Cambridge University Press.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5–8. https://doi.org/10.1111/j.1745-3992.1997.tb00585.x

Shin, S.-K. (2006). Construct validity of listening test items: A verbal protocol study. *English Teaching*, *61*(3).

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, *24*(1), 99–128. https://doi.org/10.1177/0265532207071513

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, *1*, 147–170.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In *The concept of validity: Revisions, new directions, and applications* (pp. 19–37). Charlotte, NC: Information Age Publishing.

Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 226–235. https://doi.org/10.1080/0969594X.2015.1072084

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly*, *16*(1), 32. https://doi.org/10.2307/747348

Stauffer, R. G. (1967). Reading as a cognitive process. *Elementary English*, *44*(4), 342–348. Retrieved from http://www.jstor.org/stable/41386162

Stricker, L. J., & Attali, Y. (2014). Test takers' attitudes about the TOEFL. *ETS Research Report Series*, *2010*(1), i-16. https://doi.org/10.1002/j.2333-8504.2010.tb02209.x

Swales, J. (2004). Academic writing for graduate students: Essential tasks and skills. Retrieved from http://www.tesl-ej.org/wordpress/issues/volume8/ej32/ej32r1/?wscr.

Takala, S. (2010). Putting the CEFR to good use: Activities and outcomes in Finland. In J. Mader & Z. Urkun (Eds.), *Putting the CEFR to Good Use* (pp. 96–105). Barcelona: IATEFL. Retrieved from www.iatefl.org

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, *29*, 21–36. https://doi.org/doi:10.1017/S0267190509090035

Taylor, L. (2011). *Examining speaking : research and practice in assessing second language speaking / edited by Lynda Taylor*. *Studies in language testing: 30*. Cambridge, UK ; New York : Cambridge University Press, 2011.

Taylor, L. (2014). *A report on the review of test specifications for the reading and listening papers of the Test of English for Academic Purposes (TEAP) for Japanese university entrants*.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed method research*. Thousand Oakes: Sage Publications Inc.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin*, *104*, 385–395. https://doi.org/10.1037/0033-2909.104.3.385

Tracey, D. H., & Morrow, L. M. (2012). *Lenses on reading: An introduction to theories and models* (Second Edi). New York: Guilford Press.

Treiman, R. (2017). Linguistics and reading. In M. Aronoff & J. Rees-Miller (Eds.), *Blackwell handbook of linguistics* (2nd Ed., pp. 617–626). Cornwall: Blackwell Publishing Ltd. https://doi.org/10.1002/9780470756409

Tully, M. P. (2014). Research: Articulating questions, generating hypotheses, and choosing study designs. *The Canadian Journal of Hospital Pharmacy*, *67*(1), 31–34. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3952905/

Unaldi, A. (2004). *Construct validation of the reading subskills of the Boğaziçi University English Proficiency Test.* (Unpublished doctoral dissertation). Boğaziçi University, Istanbul, Turkey.

Unaldi, A. (2010). *Investigating reading for academic purposes: sentence, text and multiple texts*. University of Bedfordshire.

Urquhart, A. H., & Weir, C. J. (1998). *Reading in a second language : Process, product and practice*. Essex: Longman.

Vahapassi, A. (1982). On the specification of the domain of school writing. In A. C. Purves & S. Takala (Eds.), *An international perspective on the evaluation of written composition* (pp. 265–289). Oxford: Pergamon.

van der Walt, J., & Steyn, H. S. (Jr. . (2008). The validation of language tests. *Stellenbosch Papers in Linguistics*, *38*, 191–204. https://doi.org/10.5774/38-0-29

Van Ek, J., & Trim, J. (2001). *Vantage*. Cambridge: Cambridge University Press.

Verhoeven, L., Reitsma, P., & Siegel, L. S. (2011, April). Cognitive and linguistic factors in reading acquisition. *Reading and Writing*, *24*(4), 387–394. https://doi.org/10.1007/s11145-010-9232-4

Weigle, S. C. (2014). Validation of automated scores of TOEFL IBT® tasks against nontest indicators of writing ability. *ETS Research Report Series*, *2011*(2), i-63. https://doi.org/10.1002/j.2333-8504.2011.tb02260.x

Weir, C. J. (1983). *Identifying the language problems of overseas students in tertiary education in the United Kingdom (Doctoral Dissertation)*. University of London, London.

Weir, C. J. (2005a). *Language testing and validation*. New York: Palgrave Macmillan.

Weir, C. J. (2005b). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*. https://doi.org/10.1191/0265532205lt309oa

Weir, C. J. (2013). An overview of the influences on English language testing. In C. J. Weir, I. Vidakovic, & E. Galaczi (Eds.), *Measured constructs: A history of Cambridge English language examinations1913 - 2012* (pp. 1–102). Cambridge: Cambridge University Press.

Weir, C. J., Hawkey, R., Green, A., & Devi, S. (2009). *The cognitive processes underlying the academic reading construct as measured by IELTS*. *IELTS Research Reports Volume 9* (Vol. 9). Bedfordshire.

Weir, C. J., Huizhong, Y., & Yan, J. (2000). *An empirical investigation of the componentiality of L2 reading in English for academic purposes*. Cambridge: Cambridge University Press.

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

Weir, C. J., & Khalifa, H. (2008b). Applying a cognitive processing model to Main Suite Reading papers. *Cambridge ESOL: Research Notes*, (31), 11–16.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594.

Wilson, K. M. (1999). Validating a test designed to assess ESL proficiency at lower developmental levels. *English*, (September). Retrieved from http://www.ets.org/Media/Research/pdf/RR-99-23.pdf

Wu, R. Y.-F. (2011). Establishing the validity of the General English Proficiency Test reading component through a criticial evaluation on alignment with the Common European Framework of Reference. University of Bedfordshire. Retrieved from http://uobrep.openrepository.com/uobrep/handle/10547/223000

Yanagawa, K. (2012). *A partial validation of the contextual validity of the Centre listening test in Japan*. University of Bedfordshire.

Yapar, T. (2003). *Study of the predictive validity of the Başkent University English proficiency exam through the use of the two-parameter IRT model`s ability estimates*. Middle East Technical University.

Yeğin, O. P. (2003). *The Predictive validity of Başkent University proficiency exam (buepe) through the use of the three-parameter IRT model`s ability estimates*. Middle East Technical University.

Zeevat, H., Grimm, S., Hogeweg, L., Lestrade, S., & Smith, E. A. (2017). Representing the lexicon: Identifying meaning in use via underspecification. In *Proceedings of Workshop Bridging Formal and Conceptual Semantics (BRIDGE-14)*. Düsseldorf: Düsseldorf University Press.

APPENDICES

APPENDIX A: RETROSPECTIVE PROTOCOL FORM

RETROSPECTION FORM

PART A
Please fill in the blanks with your personal information.

1	Year of birth	
2	Native language(s)	
3	Languages you speak other than English	
4	Gender	1.[ ] Female 2.[ ] Male
5	Current group at the DBE	1. [ ] PIN 2.[ ] INT 3.[ ] UIN 4.[ ] ADV 5.[ ] REP 6.[ ] PILOT
6	5th MT Grade	
7	I've received English language education at high school.	1.[ ] Yes 2.[ ] No
8	I've learned reading strategies at high school.	1.[ ] Yes 2.[ ] No

Instructions:

You have 30 minutes to do the test and fill out the questionnaire.

Please mark your answers to the questions on the question sheet. After answering each question, please fill out the questionnaire for that question.

- In PART B, indicate what you did before you read the test questions.
- In PART C, tick the sentences that describe what you did when you answered each question on the test.
- In PART D, indicate how you found the answer to each question.

PART B: Tick the sentence that best describes what you did before reading the questions.

Before reading the questions, I …	Q 1	Q 2	Q 3	Q 4	Q 5	Q 6	Q 7	Q 8
1. read the text or part of it slowly and carefully								
2. read the text or part of it quickly and selectively to get a general idea of what it was about								
3. did not read the text								

PART C: Tick any sentence that describes what you did when you answered each question on the test. You may tick more than one sentence for each question on the test.

To find the answer to the question, I tried to…	Q 1	Q 2	Q 3	Q 4	Q 5	Q 6	Q 7	Q 8
1. Match words that appeared in the question with exactly the same words in the text								
2. quickly match words that appeared in the question with similar or related words in the text								
3. look for parts of the text that the writer indicates to be important								
4. read key parts of the text such as the introduction and conclusion								
5. work out the meaning of a difficult word in the question								
6. work out the meaning of a difficult word in the text								
7. use my knowledge of vocabulary								
8. use my knowledge of grammar								
9. read the text or part of it slowly and carefully								
10. read relevant parts of the text again								
11. Use my knowledge of how texts like this are organised								

PART D: Tick the sentence that describes how you found the answer to each question.

I found the answer …	Q 1	Q 2	Q 3	Q 4	Q 5	Q 6	Q 7	Q 8
1. within a single sentence								
2. by putting information together across sentences								
3. by understanding how information in the whole tex t fits together								
4. I knew the answer without reading the tex t								
5. I could not answer the question								

GERİYE DÖNÜK İNCELEME

BÖLÜM A
Lütfen aşağıdaki boşluklara kişisel bilgilerinizi giriniz.

1	Doğum yılı	
2	Ana dil(ler)	
3	İngilizce dışında bildiğiniz diğer dil(ler)	
4	Cinsiyet	1.[ ] Kadın 2.[ ] Erkek
5	Bulunduğunuz grup	1. [ ] PIN 2.[ ] INT 3.[ ] UIN 4.[ ] ADV 5.[ ] REP 6.[ ] PILOT
6	5. Ara sınav (mid-term) notunuz	
7	Lisede İngilizce eğitimi aldım.	1.[ ] Evet 2.[ ] Hayır
8	Lisede İngilizce okuma stratejilerini öğrendim.	1.[ ] Evet 2.[ ] Hayır

Yönerge:

Bu çalışmayı tamamlamak için 30 dakikanız var.

Lütfen, okuma sınavındaki her bir soruyu cevapladıktan sonra o soruyla ilgili arka sayfada bulunan B, C ve D bölümlerini doldurun.

- Bölüm B'de sorulara cevap vermeden önce metni nasıl okuduğunuzu
- Bölüm C'de her bir soruyu cevaplarken neler yaptığınızı,
- Bölüm D'de her soru için cevapları nereden bulduğunuzla ilgili cümleler arasında seçim yapınız. |

BÖLÜM B: Sınav sorularını okumadan önce yaptıklarınızı tanımlayan cümlenin yanındaki kutuyu işaretleyin.

Soruları okumadan önce ...	**Soru 1**	**Soru 2**	**Soru 3**	**Soru 4**	**Soru 5**	**Soru 6**	**Soru 7**	**Soru 8**
1. metnin tümünü veya bir bölümünü yavaşça ve dikkatlice okudum.								
2. metnin tümünü veya bir bölümünü hızlıca ve seçerek genel bir fikir almak için okudum.								
3. metni okumadım.								

BÖLÜM C: Sınavdaki her bir soruyu cevaplamak için yaptıklarınızı tanımlayan cümleleri işaretleyin. Her soru için birden fazla cümle işaretleyebilirsiniz.

Sorunun cevabını bulabilmek için ...	**Soru 1**	**Soru 2**	**Soru 3**	**Soru 4**	**Soru 5**	**Soru 6**	**Soru 7**	**Soru 8**
1. soruda geçen kelimelerin aynısını metinde bularak eşleştirdim.								
2. soruda geçen kelimelerin benzerini veya ilgili olabilecek kelimeleri metinde bularak hızlıca eşleştirdim.								
3. metnin, yazarın önemli olduğunu belirttiği bölümlerine baktım.								
4. metnin giriş veya sonuç gibi önemli bölümlerini okudum.								
5. sorudaki zor bir kelimenin anlamını bulmaya çalıştım.								
6. kelime bilgimi kullandım.								
7. gramer bilgimi kullandım.								
8. metnin tümünü veya bir bölümünü yavaşça ve dikkatlice okudum.								
9. metnin ilgili bölümünü tekrar okudum.								
10. bu tür metinlerin nasıl düzenlendiğine dair bilgimi kullandım.								
11. metindeki bilgiyle kendi bilgimi birleştirdim.								

BÖLÜM D: Doğru olduğunu düşündüğünüz cevabı nasıl bulduğunuzu belirten cümleyi işaretleyin

Cevabı ...	**Soru 1**	**Soru 2**	**Soru 3**	**Soru 4**	**Soru 5**	**Soru 6**	**Soru 7**	**Soru 8**
1. tek bir cümlede buldum.								
2. birkaç cümleden aldığım bilgiyi birleştirerek buldum.								
3. metnin bütününde bilgilerin birbirine nasıl bağlandığını anlayarak buldum.								
4. metni okumadan cevabı biliyordum.								
5. soruya cevap veremedim.								

APPENDIX B: READING TEST V1 PILOT ADMINISTRATION INSTRUCTIONS

Procedure

1. **Explain the study.**
   We are doing research on a reading test that is relevant to our department. The study will take about 30 minutes and it involves answering a reading test and filling a form while doing so. We are looking for volunteers to take part in this. There is no limit to the number of the participants; we would like all of you to take part in it.
2. **Distribute the consent forms.**
   Collect them after the students read and signed them.
3. Tell the students to answer this test to the best of their ability and **ask them to change into exam position**.
4. **Distribute the protocol forms.**
   Ask students to fill in the first part. When everyone finishes filling in the personal information, explain parts B, C and D.
5. **Distribute exam papers,** and record the starting time on Test Administration Report.
6. We are expecting th estudents to finish in about 20-25 minutes. But, if there's any student who needs more time, can have time.
7. Record the finishing time on the first student handing in the exam paper.
8. Record the finishing time on the last student handing in the exam paper.
9. On **Test Administration Report** recod the time span when the majority of the students handed in their papers.
10. During the exam, if you receive any questions regarding the procedure, exam questions or the protocol form, answer them, and record them on the **Test Administration Report.**
11. Wait till the class bell rings.
12. Put all exam papers and the report in the envelope and seal it. Make sure stduents do not take copies of the exam using cell phones or other means.
13. Tell the students that they will not receive a grade from this exam but they can learn how many questions they have answered correctly. For this, after May 23, they need to call Zeynep Akşit (3952) or drop by her office in D Building.

OKUMA SINAVI DENEMESİ
UYGULAMA YÖNERGESİ

Uygulanacak İşlemler

1. **Konuyu açıklayın**:
   Bölümümüzde yapılan sınavlarla ilgili bir araştırma yapıyoruz. Bunun için sizlerle 30 dakika kadar sürecek bir çalışma yürütmek istiyoruz. Bu çalışma bir okuma sınavının çözülmesi ve bu esnada yaptıklarınızla ilgili bazı soruların cevaplanmasını kapsıyor. Bu, gönüllü katılım gerektiren bir çalışmadır. Hepinizin katılımı bizim için önemli.
2. **Gönüllü katılım formunu dağıtın.** Okunup imzalandıktan sonra formları toplayın.
3. Bu sınavı gerçek bir sınav alır gibi, yapabilecekleri en iyi şekilde yapmalarını istediğinizi söyleyin. **Öğrencileri sınav oturuşuna geçirin.**
4. **Anket formunu dağıtın.** Ön kısmını (kişisel bilgiler) doldurmalarını isteyin. Daha sonra formun arka kısmındaki üç ayrı bölümü ayrı ayrı açıklayın.
5. **Sınav kağıtlarını dağıtıp saati Test Administration Report kağıdında *Exam Start* alanına kaydedin.**
6. Ortalama 20-25 dakika içinde bitmesini bekliyoruz ama daha çok zamana ihtiyaç duyan olursa bekleyin.
7. İlk bitiren öğrencinin sınav kağıdı üzerine saati kaydedin.
8. Son bitiren öğrencinin sınav kağıdı üzerine saati kaydedin.
9. Çoğunluk öğrencinin sınavı bitirdiği süreyi **Test Administration Report** kağıdında *Exam End* alanına kaydedin.
10. Sınav sırasında gelen, uygulama şekliyle ilgili tüm soruları yanıtlayın ve arka sayfadaki tutanağa kaydedin.
11. Ders çıkış zili çalana kadar bekleyin.
12. Sınava ve araştırmaya ait (boş ve dolu) tüm kağıtları zarfa koyup kapatın. Öğrencilerin bu sınav sorularını cep telefonu vb yöntemle almamaları için özen göstermek gerekir.
13. Bu uygulama ile ilgili öğrencilere not bildirimi yapılmayacak. Ancak kaç soruyu doğru yaptığını öğrenmek isteyen olursa, 23 Mayıs'tan itibaren Zeynep Akşit'i arayıp (3952) veya ofisine giderek öğrenebilir.

APPENDIX C: READING TEST V1 PILOT ADMINISTRATION REPORT

Class:			
Proctor:			
Exam start:		End:	
Questions / Comments			

APPENDIX D: THINK ALOUD TRAINIG

Think Aloud Training

The brain is very active when taking a test. That is the reason you feel tired after a reading test: it is a difficult work because you try to make sense of what you read and decide how to answer the questions. Normally, you do not talk during reading exams. That's why no one knows what you think or do to answer the questions.

In this study, I want to learn what you think and do to find the answers to the reading questions. So, when you are answering the questions I want you to think aloud and talk out loud every thought and tell me everything you do. What I mean is, when you mark an option from among multiple options in a question, I want to understand the processes you go through: how you decide to chose that option, and/or how you eliminate the other options. Similarly, when answering short answer questions, talk out loud all the processes that you go through while trying to figure out what you think is the correct answer. At first, you may feel odd talking to yourself but it would be easier if you imagine yourself alone in the room. Once you start, you will feel more comfortable in thinking aloud.

It is important that you talk as much as possible. The aim here is to understand all the processes you go through while answering the questions. I can only do that if you say them out loud. If you remain silent for more than a few seconds, I will interrupt and ask you what you are thinking. Your thoughts may be in Turkish or in English. You can talk in whichever language you think. If you think in English, talk in English; if you think in Turkish, talk in Turkish.

While taking the test, I want to understand not only what you do but also why you do that. If, during test taking, you do not explain the reason for doing something, I may ask you to explain. For example, if you look at a question and options, and say, 'a' is the correct answer, I may ask you how you decide on that option. After a short trial, you will understand what I mean. Before I demonstrate how to think aloud while answering a reading question, I will repeat important points:

- read the question aloud and think aloud while trying to find the answer
- if your thoughts are in English, talk in English; if your thoughts are in Turkish, talk in Turkish

- if you keep silent for a few seconds, I will interrupt and ask you what you are thinking,
- I may ask you questions if I need more information on your thoughts
- I cannot help you answer the questions
- while you try this procedure with the sample question, I will start the recorder so you'll get used to its presence.

Sesli Düşünme Eğitimi

Bir okuma sınavı yaparken beyin çok aktiftir. Bu yüzden sınavlar sırasında yoruluruz, zor bir çalışmadır: Okuduğunu anlaman, ve sorulara nasıl cevap vereceğine karar vermen gerekir. Okuma sınavlarında genellikle konuşmayız, o yüzden bir soruya cevap verirken ne düşündüğünü veya yaptığını kimse bilmez.

Bu çalışmada her bir okuma sınav sorusu üzerinde çalışırken ne düşündüğünü ve ne yaptığını öğrenmek istiyorum. Senden sesli düşünmeni ve sorulara cevap verirken düşündüğün ve yaptığın her şeyi tarif etmeni istiyorum. Yani, çoktan seçmeli sorularda bir şıkkı doğru olduğunu düşünerek işaretlediğinde nasıl akıl yürüttüğünü ve/veya diğer şıkları nasıl elediğini bilmek istiyorum. Aynı şekilde, açık uçlu sorulara cevap verirken, doğru olduğunu düşündüğün cevabı bulmak için geçtiğin tüm aşamaları sesli konuşmanı istiyorum. İlk başta kendi kendine konuşmak sana garip gelebilir ama odada yalnız olduğunu hayal edersen daha kolay olur. Başladıktan sonra düşündüğünü sesli konuşmayı daha rahat yapacaksın.

Mümkün olduğunca çok konuşman çok önemli. Bu çalışmayı her soruya cevap bulurken ne düşünüp yaptığını anlamak için yürütüyoruz. Bunu da ancak sen her düşündüğünü sesli söylersen yapabiliriz. Eğer birkaç saniyeden fazla sessiz kalırsan, konuşman gerektiğini hatırlatmak için sana “ne düşünüyorsun” diye soracağım. Düşüncelerin hem Türkçe hem İngilizce olabilir. Hangi dilde düşünüyorsun o dilde konuş. Eğer İngilizce düşünüyorsan düşüncelerini İngilizce söyle; Türkçe düşünüyorsan Türkçe söyle.

Sınavı alırken sadece ne yaptığını değil, neden öyle yaptığını da anlamak istiyoruz. Sen sesli düşünürken, yaptığın şeyi neden yaptığını anlayamazsam açıklamanı isteyebilirim. Örneğin, bir sınav sorusuna bakıp doğru cevap A dersen, neden böyle karar verdin diye sorabilirim. Biraz deneme yapınca senden yapmanı istediğim şeyi daha iyi anlayacaksın. Ben senin için sesli düşünerek bir sınav sorusu okuma ve cevaplamayı örnek olarak yapacağım. Önemli noktaları kısaca tekrarlıyorum:

- soruyu okurken ve cevabı bulurken düşündüğün her şeyi sesli düşün
- düşüncelerin İngilizce ise İngilizce, Türkçe ise Türkçe konuş,
- birkaç saniye sessiz kalırsan sana “ne düşünüyorsun” diye soracağım
- düşüncelerin hakkında daha çok bilgi almak istersem sana soru sorabilirim
- soruları çözmende yardımcı olamam
- sen deneme yaparken ben de kayıt cihazını deneyeceğim, böylece buna da alışabilirsin.

APPENDIX E: READING TEST SPECIFICATIONS

READING TEST SPECIFICATIONS	
General purpose of the test	To evaluate English reading ability of non-native students in order to decide whether it is at a satisfactory level to carry out reading requirements in academic programs.
Related TLU task	Background reading to prepare for class Reading to extract important information to prepare for exams, projects Reading to understand instructions
Test taker profile	The test takers are students who have finished their secondary education and newly registered to an academic degree program at a university. Their average age is 19. They are mostly Turkish nationals.
Test level	The difficulty level of the test is set to B2[5] in the Common European Framework of Reference (Council of Europe, 2009).
Test construct	The test is designed for selection purposes. The abilities measured in this test are given in 'Careful Reading Skills Operations' and 'Expeditious Reading Skills Operations' tables below.
Test format	The test has two parts and a total of 30 items. Part 1 is Careful Reading with three tasks and 22 items. Part 2 is Expeditious Reading with one task and 8 items.
Length and administration	Test duration is 70 minutes. The test is administered in two stages. First, booklets with careful reading tasks are distributed. After 45 minutes, careful reading booklets are collected and without any break, expeditious reading task is distributed, to be completed in 25 minutes.
Characteristics of expected response	Careful Reading task items are all in the selected response format: Multiple choice with 3-options and matching. Expeditious Reading task items are in the constructed response format: Short answer.

[5] Although the test was aimed at this level, whether the items do represent the B2 level has not yet been researched; hence, it is hypothesized at this time that the test is at B2 level.

CAREFUL READING TASK SPECIFICATIONS	
Skill	Reading texts carefully to - understand the main ideas and important details - follow the development of arguments - make inferences and draw conclusions - understand the writer's organization of the text.
Time given	45 minutes for 3 texts
Task type	Reading texts and choosing the best option that answers the questions about the texts.
Instructions for the test takers	There are three texts below. Each text is followed by questions about it. For each multiple choice question, choose the best answer from among the given options, and mark your answer on the answer sheet. For each matching question, match the options with names.
Expected response	Selected response: 3-option multiple choice and matching
Items per part	6 or 7 discrete items per text

Continued **CAREFUL READING TASK SPECIFICATIONS**		
Input text: Contextual parameters	**Nature of texts**	Contemporary texts written for a non-specialist audience
	Background knowledge	Test takers should not be able to answer the questions with background knowledge without recourse to the text.
	Source of texts	Journal/newspaper articles, book sections, abstracts, text books, blogs, reports, narratives
	Text purpose	Referential, conative
	Discourse mode	Expository, argumentative, narrative
	Rhetorical organization	Explicit and implicit
	Nature of information	Concrete and abstract information at varying ratios. Pure concrete/abstract texts are not suggested
	Channel of presentation	Verbal and non-verbal (images related to the text, or graphs supporting verbal information)
	Size of input	Three texts: each text has approximately 900 words divided into paragraphs
Input text: Level	**CEFR level**	B1 to B2
	Lexical range	No technical jargon, approximately 6000 words cover 95% of the texts
	Words/sentence	15-25
	Flesch Kincaid	45-65
Speed of processing	50-70 wpm	
Task level	B1 to B2	
Topic	A broad range of topics will be selected from areas relevant to first year undergraduate students. Topics that are advised to avoid are: war, sex, religion, and fatal health issues.	
Assessment	Correct answers are scored out of 21 points.	
Weighting	All items are equally weighted.	
Marking	All items are objectively marked.	

EXPEDITIOUS READING TASK SPECIFICATIONS		
Skill	Quickly reading extended texts to - establish the gist or macrostructure of a text - find the location of a predetermined topic - find words/figures/dates	
Time given	25 minutes for one text	
Task type	Quickly reading texts and writing a response to a short answer question	
Instructions for the test takers	There a text below. It is followed by questions about it. Write a short answer that responds each question.	
Expected response	Constructed response: short answer	
Items per part	9 discrete items	
Input text: **Contextual parameters**	Nature **of texts**	Contemporary texts written for a non-specialist audience
	Background knowledge	Test takers should not be able to answer the questions with background knowledge without recourse to the text.
	Source of texts	Journal/newspaper articles, book sections, abstracts, text books, blogs, reports, narratives
	Text purpose	Referential, conative
	Discourse mode	Expository, argumentative, narrative
	Rhetorical organization	Explicit and implicit
	Nature of information	Concrete and abstract information at varying ratios. Pure concrete/abstract texts are not suggested
	Channel of presentation	Verbal and non-verbal (images related to the text, or graphs supporting verbal information)
	Size of input	One text with approximately 2500 words divided into paragraphs

Continued **EXPEDITIOUS READING TASK SPECIFICATIONS**		
Input text: Level	**CEFR level**	B1 to B2
	Lexical range	No technical jargon, approximately 6000 words cover 905 of the texts
	Words per sentence	15-25
	Flesch Kincaid level	45-65
Speed of processing	80 – 100 wpm	
Task level	B1 to B2	
Topic	A broad range of topics will be selected from areas relevant to first year undergraduate students. Topics that are advised to avoid are: war, sex, religion, and fatal health issues.	
Assessment	Correct answers are scored out of 9 points.	
Weighting	All items are equally weighted.	
Marking	All items are clerically marked according to the answer key and revised after an overview of about 10-15% of test takers' responses. In principle, grammar, punctuation, and capitalization mistakes are not penalized as long as they do not lead to a misrepresentation of the answer/mislead the reader. No partial marking.	

CAREFUL READING SKILLS OPERATIONS

	1. Understanding a Text		2. Understanding Lexis		3. Understanding Syntax	
Purpose	- To process a text to understand the main ideas and important details - To establish a macro structure for the text		- To resolve lexical ambiguity - To predict meaning of unknown words		- To identify the constituents of a sentence to understand the meaning of a text	
Operationalization	- Separate main ideas from supporting details by recognizing topic sentences and lexical indicators - Generate a representation of a text as a whole - Understand the development of an argument and/or logical organization - Make propositional inferences by using explicit statements to come to a conclusion - Distinguish generalizations and examples		- Choose between two or more meanings of a lexical item to resolve lexical ambiguity - Use contextual clues to predict the meaning of unknown words		- Remove all optional elements of a complex sentence until the sentence is left with a simple structure to understand - Paraphrase optional elements of a complex sentence one by one and fits them into the whole structure to make sense of it	
Focus	Global and local		Mainly local, sometimes global		Local	
Text coverage	Whole of text		Usually immediate context around the unknown word, sometimes a wider context is needed to identify ref.		Immediate context	
Cognitive processing for Careful Reading (Local and Global)	**Mainly bottom-up with some top-down processing**		**Mainly bottom-up processing**		**Bottom-up processing**	
	Word Recognition	√	**Word Recognition**	√	**Word Recognition**	√
	Lexical access	√	**Lexical access**	√	**Lexical access**	√
	Syntactic parsing	√	**Syntactic parsing**	√	**Syntactic parsing**	√
	Establishing propositional meaning	√	**Establishing propositional meaning**	√	**Establishing propositional meaning**	√
	Inferencing	√	**Inferencing**	√	Inferencing	
	Building a mental model	√	Building a mental model		Building a mental model	
	Creating a text level representation	√	Creating a text level representation		Creating a text level representation	
	Creating an intertextual representation		Creating an intertextual representation		Creating an intertextual representation	

EXPEDITIOUS READING SKILLS OPERATIONS

	1. **Skimming**		2. **Scanning**		3. **Search Reading**	
Purpose	- To establish the gist of the text - To establish the macro structure of the text without decoding the whole of the text - To decide the relevance of the text to established needs		- To look quickly through a text to locate a specific symbol		- To locate information on a predetermined topic	
Operationalization	- Read titles and subtitles - Read abstract - Read introductory and concluding paragraphs carefully - Read first and last sentences of each paragraph carefully - Glance at words and phrases quickly		- Match words, phrases, figures, dates, names, etc.		- Be alert for words in the same or related semantic field - Use formal knowledge of text structure for locating information - Use titles and subtitles - Read abstract where appropriate - Scan the text for words or phrases	
Focus	Global and local		Local		Global and local	
Text coverage	Selective reading		Ignores most of text		Selecting information relevant to the predetermined topic	
Cognitive processing for Expeditious Reading (Local and Global)	**Both top-down and bottom-up processing**		**Mainly bottom-up processing**		**Both top-down and bottom-up processing**	
	Word Recognition	√	**Word Recognition**	√	**Word Recognition**	√
	Lexical access	√	**Lexical access**	√	**Lexical access**	√
	Syntactic parsing	√	Syntactic parsing		**Syntactic parsing**	√
	Establishing propositional meaning	√	Establishing propositional meaning		**Establishing propositional meaning**	√
	Inferencing	√	Inferencing		**Inferencing**	√
	Building a mental model		Building a mental model		**Building a mental model**	√
	Creating a text level representation		Creating a text level representation		Creating a text level representation	
	Creating an intertextual representation		Creating an intertextual representation		Creating an intertextual representation	

Adapted from Urquhart and Weir (1998).

APPENDIX F: TEXT MAPPING FORM

Text Mapping

Date:

Text:

Mapping: Main idea, important information

EXP-M: Understand explicitly stated ideas

IMP-M: Understand implicitly stated ideas

#	TYPE	POINT	AGREEMENT

APPENDIX G: STRATEGY RUBRIC

1) Reading Strategies

A) Approaches to reading the text

RS1 Plans a goal for the text.

RS2 Makes a mental note of what is learned from the pre reading.

RS3 Considers prior knowledge of the topic.

RS4 Reads the whole text carefully.

RS5 Reads the whole text rapidly.

RS6 Reads a portion of the text carefully.

RS7 Reads a portion of the text rapidly looking for specific information.

RS8 Looks for markers of meaning in the text (e.g., definitions, examples, indicators of key ideas, and guides to paragraph development).

RS9 Repeats, paraphrases, or translates words, phrases, or sentences—or summarizes paragraphs/whole text—to aid or improve understanding.

RS10 Identifies an unknown word or phrase.

RS11 Identifies unknown sentence meaning.

RSNEW1 Skims text (reads titles, subtitles, first and last sentences of each paragraph, first and last paragraphs, takes a quick look at the text randomly reading a few words and phrases in each paragraph)

RSNEW2 Reads the whole text/paragraph one more time carefully

B) Uses of the text and the main ideas to help in understanding

RS12 During reading rereads to clarify the idea.

RS13 During reading asks self about the overall meaning of the whole text/portion.

RS14 During reading monitors understanding of the whole text/portion's discourse structure

RS15 Adjusts comprehension of the passage as more is read: Asks if previous understanding is still accurate given new information.

RS16 Adjusts comprehension of the text as more is read: Identifies the specific new information that does or does not support previous understanding.

RS17 Confirms final understanding of the text based on the content and/or the discourse structure.

C) Identification of important information and the discourse structure of the passage

RS18 Uses terms already known in building an understanding of new terms.

RS19 Identifies and learns the key words of the text.

RS20 Looks for sentences that convey the main ideas.

RS21 Uses knowledge of the whole text/portion: Notes the discourse structure of the whole text / portion (cause / effect, compare / contrast, etc.).

RS22 Uses knowledge of the whole text/portion: Notes the different parts of the text (introduction, examples, transitions, etc.) and how they interrelate.

RS23 Uses knowledge of the whole text/portion: Uses logical connectors to clarify content and passage organization.

RS24 Uses other parts of the text to help in understanding a given portion: Reads ahead to look for information that will help in understanding what has already been read.

RS25 Uses other parts of the text to help in understanding a given portion: Goes back in the text to review/understand information that may be important to the remaining text.

D) Inferences

RS26 Verifies the referent of a pronoun.

RS27 Infers the meanings of new words by using work attack skills: Internal (root words, prefixes, etc.)

RS28 Infers the meanings of new words by using work attack skills: External context (neighboring words/sentences/overall passage).

2) Test-Management Strategies

TM1 Goes back to the question for clarification: Rereads the question.

TM2 Goes back to the question for clarification: Paraphrases (or confirms) the question or task.

TM3 Goes back to the question for clarification: Wrestles with the question intent.

TM4 Reads the question and considers the options before going back to the passage/portion.

TM5 Reads the question and then reads the passage/portion to look for clues to the answer, either before or while considering options.

TM6 Predicts or produces own answer after reading the portion of the text referred to by the question.

TM7 Predicts or produces own answer after reading the question and then looks at the options (before returning to text).

TM8 Predicts or produces own answer after reading questions that require text insertion

TM9	Considers the options and identifies an option with an unknown vocabulary.
TM10	Considers the options and checks the vocabulary option in context.
TM11	Considers the options and focuses on a familiar option.
TM12	Considers the options and selects preliminary option(s) (lack of certainty indicated).
TM13	Considers the options and defines the vocabulary option.
TM14	Considers the options and paraphrases the meaning.
TM15	Considers the options and drags and considers the new sentence in context (I-it).
TM16	Considers the options and postpones consideration of the option.
TM17	Considers the options and wrestles with the option meaning.
TM18	Makes an educated guess (e.g., using background knowledge or extra-textual knowledge).
TM19	Reconsiders or double-checks the response.
TM20	Looks at the vocabulary item and locates the item in context.
TM21	Selects options through background knowledge.
TM22	Selects options through vocabulary, sentence, paragraph, or passage overall meaning
TM23	Selects options through elimination of other option(s) as unreasonable based on background knowledge.
TM24	Selects options through elimination of other option(s) as unreasonable based on paragraph/overall passage meaning.
TM25	Selects options through elimination of other option(s)has similar or overlapping and not as comprehensive.
TM26	Selects options through their discourse structure.
TM27	Discards option(s) based on background knowledge.
TM28	Discards option(s) based on vocabulary, sentence, paragraph, or passage overall meaning as well as discourse structure.
TMNEW 1	Identifies answer through keywords
TMNEW 2	Identifies answer through vocabulary, sentence, paragraph, or a number of paragraphs' overall meaning
TMNEW 3	Identifies section relevant to the question based on content

TMNEW 4	Identifies section relevant to the question: uses keywords
TMNEW 5	Identifies section relevant to the question: uses discourse structure
TMNEW 6	Identifies section relevant to the question: uses subtitles
TMNEW 7	Identifies keywords in the question
TMNEW 8	Identifies unknown vocabulary in the question
3) Test-Wiseness Strategies	
TW1	Uses the process of elimination (i.e., selecting an option even though it is not understood, out of a vague sense that the other options couldn't be correct).
TW2	Uses clues in other items to answer an item under consideration.
TW3	Selects the option because it appears to have a word or phrase from the text in it—possibly a key word.
TWNEW-1	Uses item sequence/location information as an aid to eliminate parts of text as non-relevant or to decide where to read

APPENDIX H: READING TASK PROTOTYPES

Careful Reading Task Prototype 1

A “Fail at life. Go bomb yourself.” Comments like this one, found on a CNN article about how women perceive themselves, are prevalent today across the internet, whether it’s Facebook, Reddit, or a news website. Such behavior can range from profanity and name-calling to personal attacks, sexual harassment, or hate speech. A recent Pew Internet Survey found that four out of 10 people online have been harassed online, with far more having witnessed such behavior. Trolling has become so rampant that several websites have even resorted to completely removing comments.

B Many believe that trolling is done by a small, vocal minority of sociopathic individuals. This belief has been reinforced not only in the media, but also in past research on trolling, which focused on interviewing these individuals. Some studies even showed that trolls have predisposing personal and biological traits, such as sadism and a propensity to seek excessive stimulation.

C But what if all trolls aren’t born trolls? What if they are ordinary people like you and me? In our research, we found that people can be influenced to troll others under the right circumstances in an online community. By analyzing 16 million comments made on CNN.com and conducting an online controlled experiment, we identified two key factors that can lead ordinary people to troll.

D We recruited 667 participants through an online crowdsourcing platform and asked them to first take a quiz, then read an article and engage in discussion. Every participant saw the same article, but some were given a discussion that had started with comments by trolls, whereas others saw neutral comments instead. Here, trolling was defined using standard community guidelines—for example, name-calling, profanity, racism, or harassment. The quiz given beforehand was also varied, to be either easy or difficult.

E Our analysis of comments on CNN.com helped to verify and extend these experimental observations. The first factor that seems to influence trolling is a person’s mood. In our experiment, people put into negative moods were much more likely to start trolling. We also discovered that trolling ebbs and flows with the time of day and day of the week, in sync with natural human mood patterns. Trolling is most frequent late at night, and least frequent in the morning. Trolling also peaks on Monday, at the beginning of the workweek. Moreover, we discovered that a negative mood can persist beyond the events that brought about those feelings. Suppose that a person participates in a discussion where other people wrote troll comments. If that person goes on to participate in an unrelated discussion, he or she is more likely to troll in that discussion too.

F The second factor is the context of a discussion. If a discussion begins with a “troll comment”, then it is twice as likely to be trolled by other participants later on, compared to a discussion that does not start with a troll comment. In fact, these troll comments can add up. The more troll comments in a discussion, the more likely that future participants will also troll the discussion. Altogether, these results show how the initial comments in a discussion set a strong, lasting precedent for later trolling.

G We wondered if, by using these two factors, we could predict when trolling would occur. Using machine-learning algorithms, we were able to forecast about 80 percent of the time whether a person was going to troll or not. Interestingly, mood and discussion context were together a much stronger indicator of trolling than identifying specific individuals as trolls. In other words, trolling is caused more by the person's environment than any inherent trait. Since trolling is situational, and ordinary people can be influenced to troll, such behavior can end up spreading from person to person. A single troll comment in a discussion—perhaps written by a person who woke up on the wrong side of the bed—can lead to worse moods among other participants, and even more troll comments elsewhere. As this negative behavior continues to propagate, trolling can end up becoming the norm in communities if left unchecked.

H Despite these sobering results, there are several ways this research can help us create better online spaces for public discussion. By understanding what leads to trolling, we can now better predict when trolling is likely to happen. This can let us identify potentially provocative discussions ahead of time and preemptively alert moderators, who can then intervene in these aggressive situations.

I Social interventions can reduce trolling. a) ■ If we allow people to remove recently posted comments, then we may be able to minimize regret from posting in the heat of the moment. Altering the context of a discussion, by prioritizing constructive comments, can increase the perception of civility. b) ■ Nonetheless, there is lots more work to be done to address trolling. c) ■ It is also important to differentiate the impact of a troll comment from the author's intent: Did the troll mean to hurt others, or was he or she just trying to express a different viewpoint? This can help separate undesirable individuals from those who just need help communicating their ideas.

J When online discussions break down, it is not just sociopaths who are to blame. We are also at fault. Many "trolls" are just people like ourselves who are having a bad day. Understanding that we are responsible for both the inspiring and depressing conversations we have online is key to having more productive online discussions.

Questions

1. **How does the information in paragraph B relate to paragraph C?**

   a) Paragraph B defines trolls, and paragraph C provides evidence that is found through text analysis on CNN.com.
   b) Paragraph B presents how trolls are generally characterized, and paragraph C opposes that view.
   c) Paragraph B presents research evidence on individual troll characteristics, and paragraph C supports it by presenting experiment results.

2. **According to the author, which factors are believed to affect trolling behavior?**

   a) Time and day, and the number of participants in a discussion
   b) People's feelings and familiarity with others they communicate with
   c) People's state of mind and interaction behavior

3. **Which of the following <u>cannot</u> be concluded from paragraph G?**

   a) Online discussion boards need to be moderated.
   b) Specific conditions accelerate trolling behavior.
   c) Computed algorithms reveal best who will troll.

4. **Choose the best summary for paragraph H.**

   a) This research is useful in revealing the reasons for trolling and preparing to take action before trolling happens.
   b) The results of the research are disheartening; however, through open discussions, we may be able to prevent trolling in online spaces such as discussion boards.
   c) The research reveals that we should be more careful in online platforms and help moderators isolate those people who troll.

5. **Where in paragraph I does the following sentence belong?**

   Even just pinning a post about a community's rules to the top of discussion pages helps, as a recent experiment conducted on Reddit showed.

   a) a
   b) b
   c) c

6. **What is the best title for this text?**

   a) Trolls redefined
   b) Trolling on the rise
   c) A troll by nature

Careful Reading Task Prototype 2

A In the 1950s, the Finnish biologist Björn Kurtén noticed something unusual in the fossilized horses he was studying. When he compared the shapes of the bones of species separated by only a few generations, he could detect lots of small but significant changes. Horse species separated by millions of years, however, showed far fewer differences in their bone structure. Subsequent studies over the next half century found similar effects—organisms appeared to evolve more quickly when biologists tracked them over shorter timescales. Then, in the mid-2000s, Simon Ho, an evolutionary biologist at the University of Sydney, encountered a similar phenomenon in the genomes he was analyzing. When he calculated how quickly DNA mutations accumulated in birds over just a few thousand years, Ho found the genomes full of small mutations. This indicated a rapidly ticking evolutionary clock. But when he zoomed out and compared DNA sequences separated by millions of years, he found something very different. The evolutionary clock had slowed to a crawl.

B Baffled by his results, Ho set to work trying to figure out what was going on. He stumbled upon Kurtén's 1959 work and realized that the differences in rates of physical change Kurtén saw also appeared in genetic sequences. His instincts as an evolutionary biologist told him that the mutation rates he was seeing in the short term were the correct ones. The genomes varied at only a few locations, and each change was as obvious as a splash of paint on a white wall. But if more splashes of paint appear on a wall, they will gradually conceal some of the original color beneath new layers. Similarly, evolution and natural selection write over the initial mutations that appear over short timescales. Over millions of years, an A in the DNA may become a T, but in the intervening time it may be a C or a G for a while. Ho believes that this mutational saturation is a major cause of what he calls the time-dependent rate phenomenon.

C "Think of it like the stock market," he said. "Look at the hourly or daily fluctuations of Standard & Poor's 500 index, and it will appear wildly unstable, swinging this way and that. Zoom out, however, and the market appears much more stable as the daily shifts start to average out. In the same way, the forces of natural selection weed out the less advantageous and more deleterious mutations over time."

D Ho's discovery of the time-dependent rate phenomenon in the genome had major implications for biologists. It meant that many of the dates they used as bookmarks when reading life's saga—everything from the first split between eukaryotes and prokaryotes billions of years ago to the re-emergence of the Ebola virus in 2014—could be wrong. "When this work came out, everyone went 'Oh. Oh, dear,'" said Rob Lanfear, an evolutionary biologist at the Australian National University in Canberra.

E The time-dependent rate phenomenon wasn't fully appreciated at first. For one thing, it is such a large and consequential concept that biologists needed time to wrap their heads around it. But there's a bigger block: The concept has been all but impossible to use. Biologists have not been able to quantify exactly how much they should change their estimates of when things happened over the course of evolutionary history. Without a concrete way to calculate the shifts in evolutionary rates over time, scientists couldn't compare dates.

F Recently, Aris Katzourakis, a paleovirologist at the University of Oxford, has taken the time-dependent rate phenomenon and applied it to the evolution of viruses. In doing so, he has not only pushed back the origin of certain classes of retroviruses to around half a billion years ago—long before the first animals moved from the seas to terra firma—he has also developed a mathematical model that can be used to account for the time-dependent rate phenomenon, providing biologists with much more accurate dates for evolutionary events.

G Other scientists are excited by the prospect. "It's like Einstein's theory of relativity, but for viruses," said Sebastián Duchêne, a computational evolutionary biologist at the University of Melbourne. The time-dependent rate phenomenon says that the speed of an organism's evolution will depend on the time frame over which the observer is looking at it. And as with relativity, researchers can now calculate by how much.

Questions

1. **What is the function of paragraph A?**

   a) It explains unexpected findings regarding the development of a rare animal species.
   b) It introduces similar research findings by two biologists from different countries.
   c) It shows how the understanding of evolutionary process varied in two decades.

2. **Why does the writer use the phrase "a splash of paint on a white wall" in paragraph B?**

   a) To help the reader recognize the significance of short-term mutation rates
   b) To help the reader see the similarity between mutation and natural selection
   c) To help the reader understand the causes of different mutation rates

3. **According to paragraph E, what is true about biologists' reactions to the time-dependent rate phenomenon?**

   a) They did not think it was such a significant find.
   b) They tried to challenge the idea with further research.
   c) They felt they needed a method to put it into practice.

4. **Which of the following could be the best title for this text?**

   a) Evolution and time: New evolutionary evidence creates a conflict
   b) DNA mutations may have been overrated, new research finds
   c) Evolution is slower than it looks, faster than you think

Match statements (11-13) with a scientist (a-e). There are more names than you need.

5. He introduced a new concept that greatly altered the existing literature of evolution. ______

6. His work enabled the putting of time-dependent rate phenomenon to practical use. ______

7. His work focused on the physical make-up of fossils belonging to an animal species. ______

a. Aris Katzourakis

b. Björn Kurtén

c. Rob Lanfear

d. Sebastián Duchêne

e. Simon Ho

Search Reading Task Prototype 1

SVALBARD SEED VAULT

I. Introduction

Mission

Though the general public is well aware of the threat of extinction to animal species, far fewer are aware of the risk of crop extinction. With whales or tigers or polar bears, you can look at them in the eye and you can be very empathetic. But you can't do that with a wheat variety or carrot variety. Preserving seed from food plants is an absolutely essential part of the work of preserving the world's biodiversity, adapting to climate change and global warming, with an eventual goal **to ensure food for the world's population**. The foundation of a global central seed bank for the world's seeds (primarily of food plants) has therefore long been an issue, and Svalbard Global Seed Vault was a step in this direction.

Funding and Construction of the Vault

The history of Svalbard seed vault starts as early as 1983. Like other big projects, it's been a long and not very easy journey. In 1989, the International Board for Plant Genetic Resources (IBPGR) started surveying the relevant alternative sites in Svalbard. Norway offered to take care of the actual construction of the vault, while the Food and Agriculture Organization (FAO) and IBPGR would take care of the administrative operational expenses through the creation of a fund based on capital from external donors.

II. Description of the facility

Location

This Seed Vault lies about 1 kilometer from Longyearbyen Airport, at about 130 meters above sea level and consists entirely of an underground facility, blasted out of the permafrost (at about minus 3-4 degrees Celsius). The facility is designed to have an almost "endless" lifetime. The location takes into account all known scenarios for rising sea level caused by global climate changes. The facility has also been located so deep inside the mountain that any possible changes to Svalbard's climate, which we know about today, will not affect the efficacy of the permafrost.

Inside the Facility

The facility consists of three separate underground chambers. The layout of these chambers is purposeful. None of them are in a direct line. Instead, the workers have carved out a concave indentation in the rock. This serves as a security measure against an explosion. The chambers, each of which with a capacity to store 1,5 million different seed samples, have storage shelving for pre-packed examples of food seeds from the depositors.

A tunnel, which is about 100 meters long, is used to access the chambers. It has an entrance portal which is the only visible part of the facility. It is in the form of a long, narrow concrete “fin”, with an entrance of brushed steel. An artistic decoration on the outer roof surface and on the upper part of the front partly reflects the polar light and partly gives off a muted, glowing light. The outer half of the entrance tunnel is constructed as a steel pipe with a diameter of about 5 meters. This passes through the layer of snow and ice and the loose rocks, into solid mountain. The innermost part and the storage chambers were blasted out of the mountain using tunnel drilling and rock blasting techniques. The mountain is secured with bolts and spray concrete. The permafrost also contributes to stability. The interior floor is of asphalt. There is electric lighting throughout and the facility is secured against forced entry and has TV surveillance. Areas for filing and other administrative work of a temporary nature are located beside the entrance tunnel. The total floor area of the facility is just less than 1,000 square meters.

III. Administration

Early Conflicts

In the early 90s, there was heated debate between the various member countries of the FAO about patenting and access to genetic resources. Developing countries wished to receive part of the proceeds from the commercial seed industry, since the diversity mainly came from their areas, whilst the commercial seed industry wanted free access to such resources and the opportunity to patent the seeds. This led to a polarized atmosphere with little mutual trust regarding the administration of seed. The lack of international agreement to regulate this area eventually became an obstacle to realizing the plans for an international safety deposit for seeds in Svalbard, and the construction of the vault had to be delayed.

Who Owns the World’s Heritage?

The turning point came when FAO’s International Treaty for Plant Genetic Resources for Food and Agriculture came into force in 2004. This created a new basis for taking the plans up again. The Norwegian Ministry of Foreign Affairs and the Ministry of Agriculture and Food took up the challenge. A group of Nordic and international experts under the direction of Noragric at the Norwegian University of Life Scientists (UMB) were appointed to carry out a preliminary study. In September 2004, the group put forward an unambiguously positive report, which concluded that suitable locations were to be found in Svalbard. The report recommended that a chamber should be built inside the mountain.

In November 2004, the report was presented at FAO's Commission for Genetic Resources for Food and Agriculture. The Norwegian idea received a positive response and was perceived by many countries as a most welcome contribution to the international work of preserving the world's plant genetic resources. Some developing countries also pointed to the earlier positive experience of development collaborations with Nordic countries and the Nordic Genetic Resource Centre in Svalbard. Following the FAO meeting Norway began work on financing the construction project. Since the purpose of the seed vault was multilateral, it was natural to pave the way for making this a joint initiative between three ministries, the Ministries of Foreign Affairs, Environment and Agriculture and Food. The government backed the initiative and in 2005 an interdepartmental steering group was set up for the project. Under the chairmanship of the Ministry of Agriculture and Food, the steering group discussed various alternatives for the location, organization, agreement format and operation of the seed vault, as well as working in close cooperation with international experts in relevant fields. It was also stressed that the storage of seeds should be done in accordance with international gene bank standards, at minus 18 degrees, and that the seeds should be stored by the black box method, which means that only the institution which deposits seeds has right of ownership and disposition over them. That is, even though the facility is owned by Norway, it is important to underline that the seed samples which are stored in the vault are indisputably the property of the depositor.

IV. Why Svalbard?

1) Svalbard, as Norwegian territory, enjoys security and political and social stability. Norway understands the importance of preserving Svalbard as an area of undisturbed nature, which is now an important research and reference area. The seed vault fits ideally into this concept.

2) Svalbard has an isolated position far out in the ocean, between 74° and 81° N and only 1000 kilometers from the North Pole. This archipelago has an undisturbed nature. Permafrost, which is characteristic of this area, provides stable storage conditions for seeds, even when there is a power cut or a technical problem with cooling systems.

3) The seed vault, which consists of three chambers, is located right outside Longyearbyen and directly opposite Longyear Airport. The facility is about 130 meters above sea level and has been tunneled 120 meters into the mountain, in a stable sandstone situation. Each of the three underground chambers is about 1,200 cubic meters (20 meters deep, 10 meters wide and 6 meters high). The location so far below ground guarantees stable permafrost for the foreseeable future and is high enough above sea level to secure the facility against any rise in sea level as a result of global warming.

4) The facility's open location near the town makes monitoring and security easier. Security is the responsibility of the Governor of Svalbard in cooperation with the University of Svalbard (UNIS).

International Significance

The international seed vault is a unique contribution to the preservation of the planet's most important biodiversity. This has been a priority issue for Norway for many years as well as the principal objective of the Biodiversity Convention and the FAO treaty. The seed vault could come to have a special significance for a number of regions in developing countries where the storage conditions in regular gene banks are a constant challenge. For many years it has been Norway's aim to play a bridge-building role in the north-south debate about genetic resources and biological diversity. Svalbard Global Seed Vault can be a unifying initiative, which offers much to countries both north and south and which will hopefully also promote global collaboration in taking care of our most important genetic resources. Securing food supplies is one of the most basic issues in any strategy for eliminating poverty. In a time of climate change, this is an equally global issue. The establishment of a global seed vault is therefore very much in line with the principle of informed self-interest.

V. Seed Storage

The Svalbard Global Seed Vault provides facilities free of cost for safety deposits under "black box conditions" on request from public or private holders of seeds of distinct genetic resources that are important to humanity. Priority is given to the safety deposit of plant genetic resources of importance for food security and sustainable agriculture.

Costs pertaining to the packaging and shipping of the deposited seeds are borne by the depositors. However, in the case of developing countries and international gene banks, the Global Crop Diversity Trust is funding the costs of preparing, packing and shipping their seeds to Svalbard.

The Seed Vault does not have the opportunity to test the viability of the seeds, but accepts new shipments of seeds when the duplicate samples at the depositor's possession have lost fertility. Import and storage of GMO (Genetically-modified) seeds according to Norwegian legislation require advance approval. Certain other criteria apply to "sealed internal use" for research purposes and indoor storage of GMO, for example with regard to the risk of spreading GMO. Norwegian gene technology legislation was formulated before the Svalbard Global Seed Vault (SGSV) was set up, and therefore fails to take into account the vault's special status, or the low risk related to handling seeds in sealed packaging. Until changes can be made to the rules, long-term storage of GMO seeds in the SGSV will not be approved.

VI. Conclusion

Svalbard Global Seed Vault is not a gene bank, it is a facility for maintaining crop diversity in the form of seeds, stored and conserved in a frozen state. The ideal temperature is between minus 10 and minus 20 degrees Celsius. Gene banks may also contain living plants and parts of plants in those cases where it is difficult to store the crop in the form of seeds. The Seeds in the Seed Vault shall only be accessed when the original seed collections have been lost for any reason.

The Seed Vault has the capacity to store 4,5 million different seed samples. Each sample contains on average 500 seeds, so a maximum of 2,25 billion seeds may be stored in the Seed Vault. The Seed Vault, therefore, has the capacity to hold all the unique seed samples that are conserved today by all the gene banks in more than 100 countries all over the world. In addition, the Seed Vault has the capacity to also store many new seed samples that may be collected in the future. When in full use, the Svalbard Global Seed Vault will represent the world's largest collection of seeds. Priority is given to crops that are important for food production and sustainable agriculture, which is of utmost importance for developing countries where food security is a challenge.

Different crop varieties have different characteristics and not all the differences may be visible to the eye. Genetic traits may provide differences in disease resistance, adaptability to various soils and climates, different tastes and nutritional qualities. If we ever need to use the potentially unique and sometimes hidden traits found in a particular crop variety, then we must ensure that the variety is available. Unfortunately, much diversity has already been lost. The number of plant varieties used during the last 30 years of intensification of agriculture has been dramatically reduced. If we do not take action immediately, different varieties of wheat and potato can disappear as permanently as dinosaurs.

Search Reading Questions

Complete the statements below using information from the text. Keep your answers as short as possible.

Sample item:

0. The ultimate purpose of founding a global central seed bank is ***to ensure food for the world's population*** **.**

1. The FAO and IBPGR agreed to cover the operation costs of the vault with the money collected from ____.

2. The specific layout of the chambers functions as ______________________________.

3. In the 1990s, there was no consensus among nations on seed administration. This led to the postponement of ____.

4. According to the black box method, the seed samples stored in the vault belong to ____.

5. The advantage of Svalbard's location in providing an appropriate setting for seed storage is ____.

6. ____ is the main aim of the Biodiversity Convention and the FAO treaty.

7. According to the laws in Norway, it will not be possible to import and store GMO seeds in Svalbard without ____.

8. ____ will benefit most from the Vault's preference for specific seed samples because their food production in the long-term is at risk.

APPENDIX I: CURRICULUM VITAE

Zeynep Akşit

zaksit@metu.edu.tr

Year of Birth: 1965

Place of Birth: Ankara, Turkey

EDUCATION

Doctoral Degree **2012 - 2018**	English Language Teaching, METU, ANKARA
Master's Degree **2009 – 2011**	Teaching English as a Foreign Language, Bilkent University, ANKARA
Bachelor's Degree **1982 – 1987**	English Language and Literature, Boğaziçi University, ISTANBUL

CAREER HISTORY

R&D Unit Coordinator **2012 – Present**	School of Foreign Languages, METU, ANKARA Job detail: Carrying out research studies in accordance with the needs of the School of Foreign Languages. Some of the projects are • Revision of METU-EPE (2015-2018) • Conducting a 30-hour training course, *Foreign Language Assessment: Theories and Practices,* at the Department of Basic English, METU (2015 – 2016) • Evaluation of the instructional program at the Department of Basic English, METU (2013 – 2015) • Analysis of METU students' communicative needs (2013-2015) • Analysis of METU-EPE (Since 2013) • Analysis of in-house achievement tests (2012-2013)

Conference Convener
2011 – 2012

Department of Basic English, METU, ANKARA
Job detail: Organized the 11th METU ELT Convention, Embracing Challenges", 31 May – 02 June 2012, METU.

Instructor
2006 - Present

Department of Basic English, METU, ANKARA

Job detail: Teaching of four skills at all levels (beginner, elementary, intermediate, upper-intermediate)

Managing Partner
1989 – 2002

Kilim Bilgi İşlem Sistemleri Ltd., ISTANBUL

Job detail: Worked as head of the training unit, and vice manager. The job involved training of the customers on using various software and hardware.

Tourist Guide
1987 - 1992

Job detail: Certified by the Ministry of Tourism to conduct tours in English. I worked as a tour guide in Turkey, and also led tours to Europe.

CONFERENCE PRESENTATIONS

Akşit, Zeynep. & Saygı, Şükran. (2017). A Multi-Method Approach to Writing Scale Development. Paper presented at the Joint Meeting & Workshop of the EALTA Special Interest Groups Assessment of Writing and Assessment for Academic Purposes. Bremen University, Bremen, Germany, 17-18 November 2017.

Saygı, Şükran & **Akşit, Zeynep.** (2016). Experiential Learning of a Test Writer. Paper presented at the 14th International Bilkent University School of English Language Conference. Bilkent University, Ankara, Turkey, 17-18 June 2016.

Akşit, Zeynep. (2016c). Beyond the Preparatory School: How Do Students Cope? Paper presented at the 9th International ELT Research Conference. Çanakkale Onsekiz Mart University, Çanakkale, Turkey, 12-14 May 2016.

Akşit, Zeynep. (2016b). Defining Academic Reading for Assessment. Paper presented at the 13th EALTA 2016 Annual Conference. Universitat Politecnica de Valencia, Valencia, Spain, 5-8 May 2016.

Akşit, Zeynep. (2016a). The State of English in Higher Education in Turkey – A Baseline Study. Invited panellist at the English and Beyond in Higher Education Conference. Istanbul Teknik University, Turkey, 15-16 Feb. 2016.

Akşit, Zeynep. (2015b). Language Assessment in Tertiary Education: The Case of Language Preparatory Schools. Paper presented at the 3rd ULEAD Congress on Applied Linguistics. Çanakkale, Turkey, 08-10 May 2015.

Akşit, Zeynep. (2015a). Academic Language Skills Needs: Are We in Accord? Paper presented at the 12th ODTÜ International ELT Convention. Department of Basic English, ODTÜ, Ankara, Turkey, 25-26 May 2015.

Akşit, Zeynep. (2014). A Validity Issue. Paper presented at the 10th Annual CamTESOL Conference on English Language Teaching. IDP Education, Phnom Penh, Cambodia, 22-23 February 2014.

Akşit, Zeynep. (2013). Attitudes Towards Research in an EFL Context. Paper presented at the 13th International ELT Conference. Bilkent University, Ankara, Turkey, 17-18 June 2013.

Akşit, Zeynep. (2011). Attitudes Towards Research in an EFL Context. Paper presented at the 45th Annual International IATEFL Conference and Exhibition. Brighton, UK, 15-19 April 2011.

Akşit, Zeynep. (2009). Using 'Track Changes' to Correct Student Writing. Paper presented at the 10th METU ELT Convention. Ankara, 22-23 May 2009.

PUBLICATIONS

Akşit, Zeynep. (2009). Using "Track Changes" in Microsoft Word® to Correct Student Writing. *Proceedings of the 10th Middle ELT Conference*. Ankara, METU.

Tekir, Serpil, **Akşit, Zeynep** & Önal, Elif. (2010). English Break (B1). Ankara, Gündüz Yayıncılık.

ASSESSMENT-RELATED TRAININGS and WORKSHOPS

December 2015	Workshop on validating CEFR descriptors for mediation activities and strategies. Council of Europe.
May 28-31, 2015	European Association for Language Testing and Assessment (EALTA) Workshop: Standard setting: How to implement good practice? (by Sauli Takala and Charalambos Kollias*)*, Copen-hagen, Denmark.

Nov 21, 2014	Focus on Assessment Issues IV Meeting. Bilkent University, Ankara.
June 2-3, 2014	Language Testing Research Colloquium (LTRC) Workshop: Using an *Assessment Use Argument* to develop language tests and justify their use (by Lyle Bachman, Adrian Palmer and Daniel Dixon), The Netherlands.
March 21, 2014	Focus on Assessment Issues III Meeting. Özyeğin University, Istanbul.
Oct 4, 2013	Focus on Assessment Issues II Meeting. Istanbul Şehir University, Istanbul.
July 28-Aug 1, 2013	EALTA Summer School: Testing and assessment for learning languages (by James Purpura, Ildiko Csepes, Gudrun Erickson, and Dina Tsagari), Università per Stranieri di Siena, Italy.
May 31, 2013	Assessment Literacy Advisory Panel Meeting II. Bilkent University, Ankara
April 8, 2013	IATEFL Leadership and Management: Professional Development Day, Liverpool, UK.
Nov 2012	Assessment Literacy Advisory Panel Meeting I. TOBB University, Ankara
Aug 6 -10, 2012	EALTA Summer School: Good practice in language testing and assessment: A psychometric perspective (by Jan-Eric Gustafsson, John de Jong, and Norman Verhelst), University of Gothenburg, Sweden.
May 17, 2011	One-day ELT Event with Dr. Krashen. Turkish Military Academy, Ankara
April 15 – 19, 2011	IATEFL Leadership and Management: Professional Development Day, Brighton, UK.

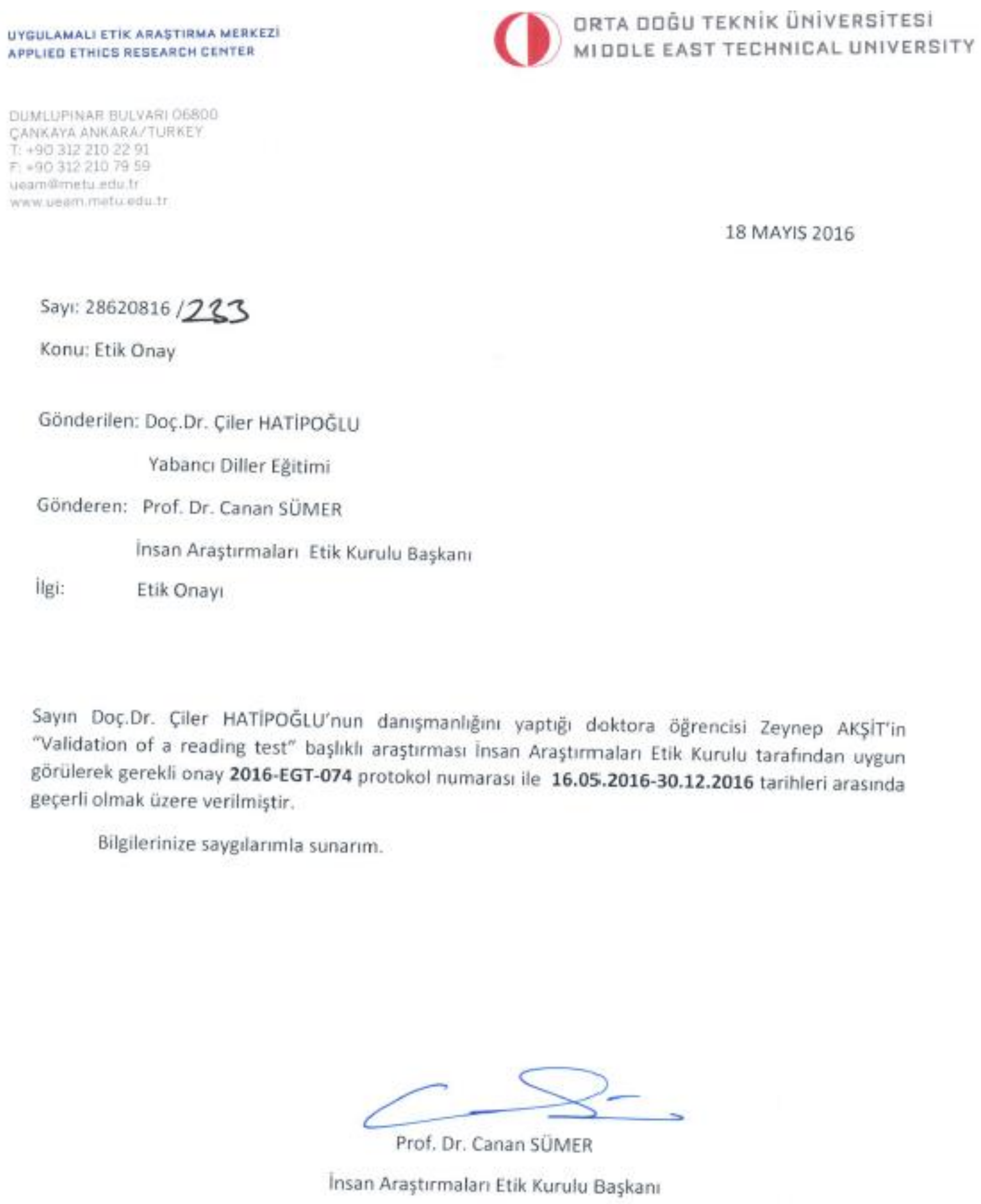
SKILLS

Languages	
Turkish	Native speaker
English	C1
French	A1/A2
Italian	A1/A2
Computer skills	Advanced

AWARDS

METU Annual Performance Award (2015 – 2017)
METU Annual Performance Award (2013 – 2015)
METU Annual Performance Award (2011 – 2013)

APPENDIX J: ETHICS COMMITTEE PERMISSION FORM



UYGULAMALI ETİK ARAŞTIRMA MERKEZİ
APPLIED ETHICS RESEARCH CENTER

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

DUMLUPINAR BULVARI 06800
ÇANKAYA ANKARA/TURKEY
T: +90 312 210 22 91
F: +90 312 210 79 59
ueam@metu.edu.tr
www.ueam.metu.edu.tr

18 MAYIS 2016

Sayı: 28620816 /233

Konu: Etik Onay

Gönderilen: Doç.Dr. Çiler HATİPOĞLU
Yabancı Diller Eğitimi

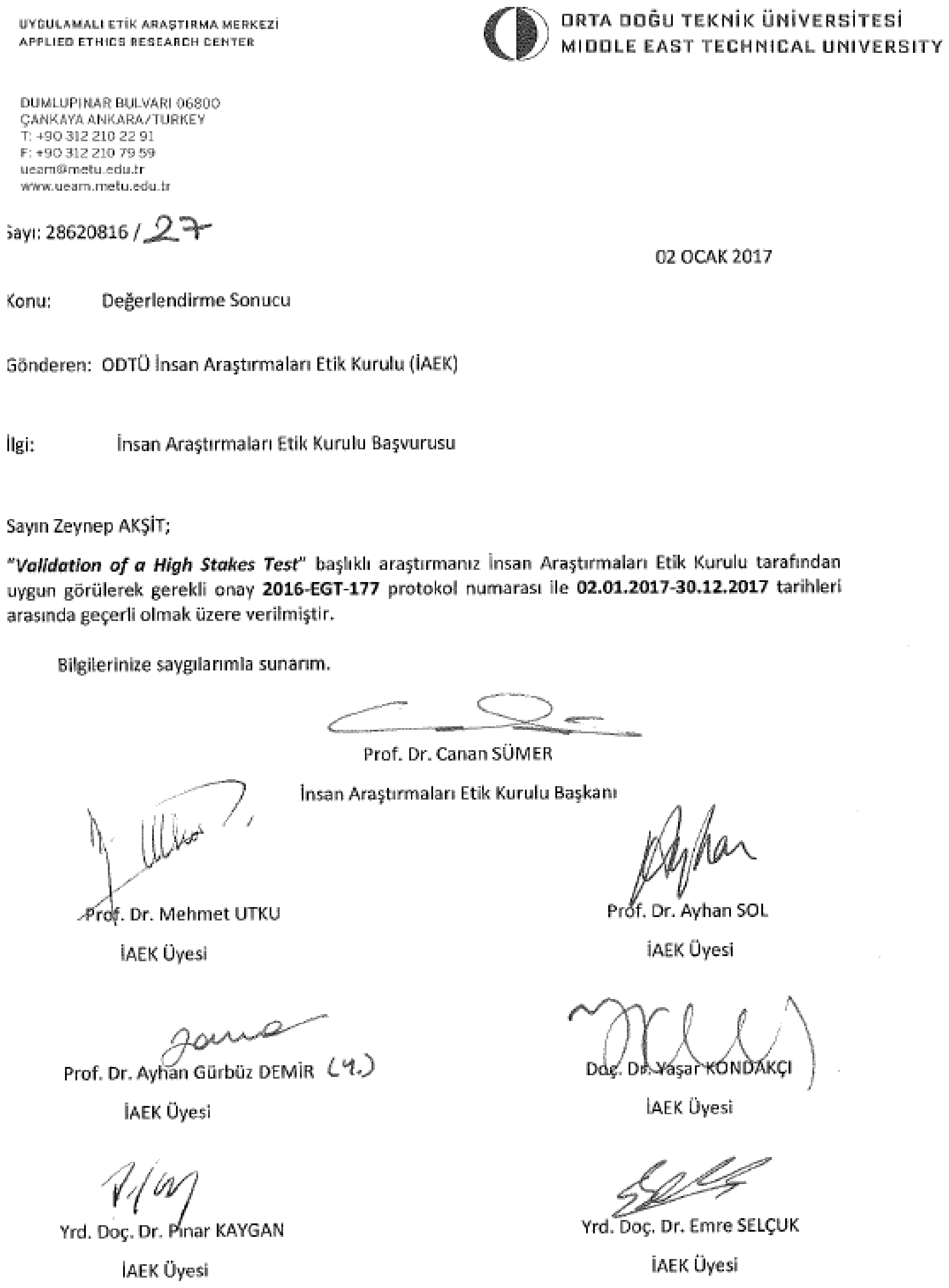
Gönderen: Prof. Dr. Canan SÜMER
İnsan Araştırmaları Etik Kurulu Başkanı

İlgi: Etik Onayı

Sayın Doç.Dr. Çiler HATİPOĞLU'nun danışmanlığını yaptığı doktora öğrencisi Zeynep AKŞİT'in "Validation of a reading test" başlıklı araştırması İnsan Araştırmaları Etik Kurulu tarafından uygun görülerek gerekli onay **2016-EGT-074** protokol numarası ile **16.05.2016-30.12.2016** tarihleri arasında geçerli olmak üzere verilmiştir.

Bilgilerinize saygılarımla sunarım.

Prof. Dr. Canan SÜMER
İnsan Araştırmaları Etik Kurulu Başkanı



UYGULAMALI ETİK ARAŞTIRMA MERKEZİ
APPLIED ETHICS RESEARCH CENTER

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

DUMLUPINAR BULVARI 06800
ÇANKAYA ANKARA/TURKEY
T: +90 312 210 22 91
F: +90 312 210 79 59
ueam@metu.edu.tr
www.ueam.metu.edu.tr

Sayı: 28620816 / 27

02 OCAK 2017

Konu: Değerlendirme Sonucu

Gönderen: ODTÜ İnsan Araştırmaları Etik Kurulu (İAEK)

İlgi: İnsan Araştırmaları Etik Kurulu Başvurusu

Sayın Zeynep AKŞİT;

"Validation of a High Stakes Test" başlıklı araştırmanız İnsan Araştırmaları Etik Kurulu tarafından uygun görülerek gerekli onay **2016-EGT-177** protokol numarası ile **02.01.2017-30.12.2017** tarihleri arasında geçerli olmak üzere verilmiştir.

Bilgilerinize saygılarımla sunarım.

Prof. Dr. Canan SÜMER
İnsan Araştırmaları Etik Kurulu Başkanı

Prof. Dr. Mehmet UTKU
İAEK Üyesi

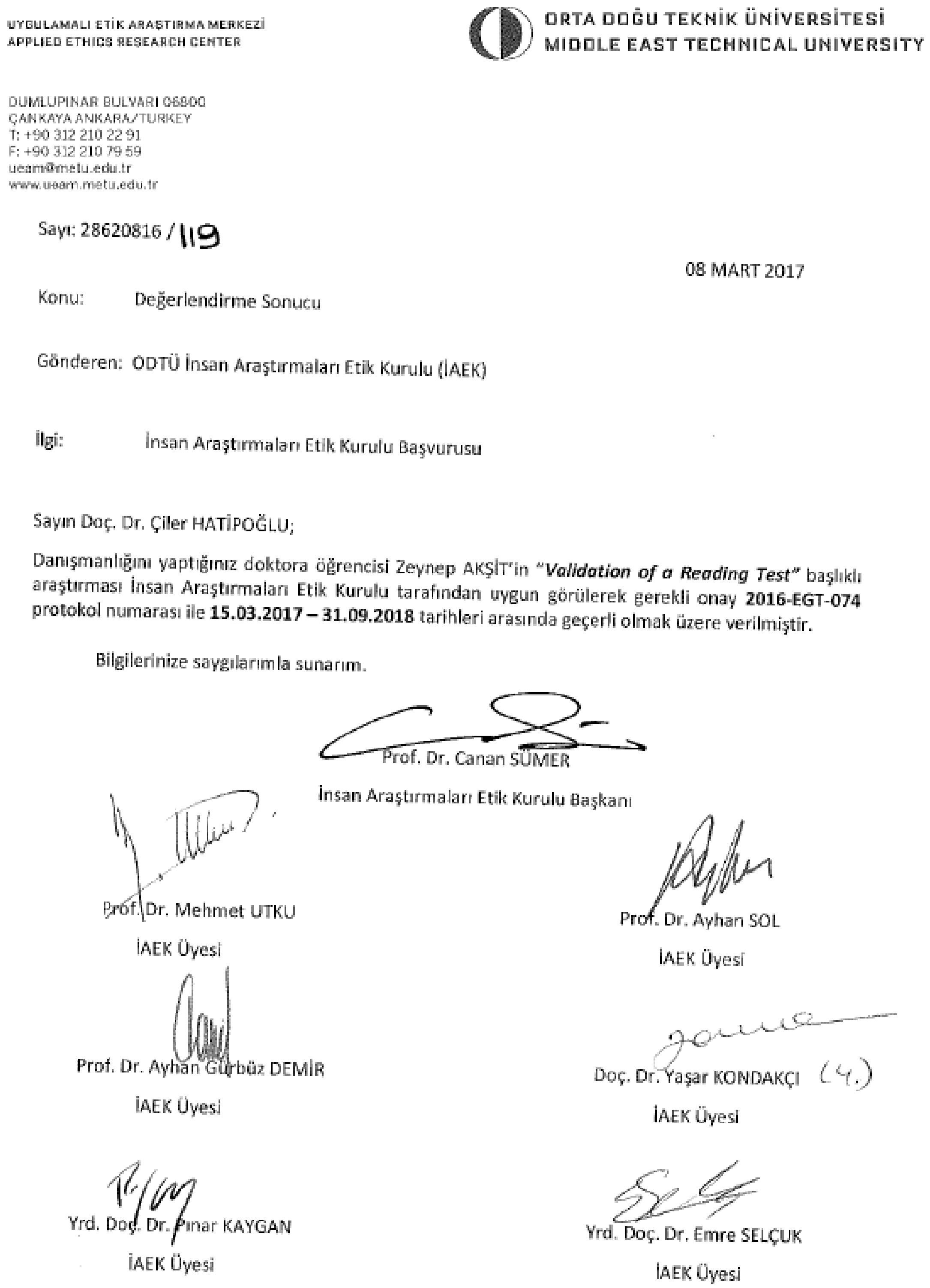
Prof. Dr. Ayhan SOL
İAEK Üyesi

Prof. Dr. Ayhan Gürbüz DEMİR (4.)
İAEK Üyesi

Doç. Dr. Yaşar KONDAKÇI
İAEK Üyesi

Yrd. Doç. Dr. Pınar KAYGAN
İAEK Üyesi

Yrd. Doç. Dr. Emre SELÇUK
İAEK Üyesi



UYGULAMALI ETİK ARAŞTIRMA MERKEZİ
APPLIED ETHICS RESEARCH CENTER

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

DUMLUPINAR BULVARI 06800
ÇANKAYA ANKARA/TURKEY
T: +90 312 210 22 91
F: +90 312 210 79 59
ueam@metu.edu.tr
www.ueam.metu.edu.tr

Sayı: 28620816 / 119

08 MART 2017

Konu: Değerlendirme Sonucu

Gönderen: ODTÜ İnsan Araştırmaları Etik Kurulu (İAEK)

İlgi: İnsan Araştırmaları Etik Kurulu Başvurusu

Sayın Doç. Dr. Çiler HATİPOĞLU;

Danışmanlığını yaptığınız doktora öğrencisi Zeynep AKŞİT'in "***Validation of a Reading Test***" başlıklı araştırması İnsan Araştırmaları Etik Kurulu tarafından uygun görülerek gerekli onay **2016-EGT-074** protokol numarası ile **15.03.2017 – 31.09.2018** tarihleri arasında geçerli olmak üzere verilmiştir.

Bilgilerinize saygılarımla sunarım.

Prof. Dr. Canan SÜMER
İnsan Araştırmaları Etik Kurulu Başkanı

Prof. Dr. Mehmet UTKU
İAEK Üyesi

Prof. Dr. Ayhan SOL
İAEK Üyesi

Prof. Dr. Ayhan Gürbüz DEMİR
İAEK Üyesi

Doç. Dr. Yaşar KONDAKÇI (4.)
İAEK Üyesi

Yrd. Doç. Dr. Pınar KAYGAN
İAEK Üyesi

Yrd. Doç. Dr. Emre SELÇUK
İAEK Üyesi

APPENDIX K: TURKISH SUMMARY / TÜRKÇE ÖZET

GİRİŞ

Yüksek eğitim kurumlarında çeşitli sebeplerle dil becerilerini ölçme sınavları uygulanır. Bu sınavların bir amacı sınavı alan adayların akademik çalışma yapabilecek düzeyde dil becerisine sahip olup olmadığını ölçmektir. Bu sebeple kullanılan dil yeterlilik sınavları uluslararası kurumlar tarafından hazırlanan sınavlar olabileceği gibi (örneğin, ETS kurumu tarafından hazırlanan TOEFL veya British Council – IDP Education işbirliği ile hazırlanan IELTS), üniversitelerin kendi bünyelerinde de hazırlanabilir. Bu sınavların sonuçları sınanan kişiler ve sınavı veren kurumlar açısından önemli sonuçlar doğurduğundan, örneğin, bir programa kabul veya red kararı, "yüksek etkili sınav" (high stakes test) olarak adlandırılırlar. Yüksek etkili sınavlarda kullanılan değerlendirme aracının adil ve doğru şekilde ölçmesini sağlamak sınav geliştiricilerinin sorumluluğundadır. Bu açıdan yüksek etkili sınavların tasarımı, geliştirmesi ve uygulanmasında temel unsurlardan biri geçerlik kavramıdır.

Geçerlik bir sınavın anlamlı, yararlı ve adil olmasıyla ilgilidir (Messick, 1989b). Günümüzde geçerlik, sınav notuna dayanarak verilen kararların ne derecede gerekçelendirilebildiğiyle ölçülür (Chapelle, 1998; Council of Europe, 2009; Cronbac, 1988; Fulcher & Davidson, 2007Kane, 2016; Mcnamara, 2006; Messick, 1995; Sireci, 2009). Yabancı dil sınav puanına dayanarak verilen bir akademik programa kabul veya red kararının araştırmalarla desteklenerek geçerliği gösterilmelidir. Ölçünleştirilmiş dil sınavları üreten pek çok kurum sınavlarının anlamlı, yararlı ve adil olduğunu göstermek üzere geniş kapsamlı araştırmalar yaparlar (Örneğin, ETS, Cambridge ESOL, British Council - IDP Education). Bu araştırmalar sınav kurgusu, içerik geçerliği, kestirim geçerliği, veya kriter geçerliği gibi konularda olabilir. Bu çalışmada, bir okuma sınavının bağlamsal, bilişsel ve notlama geçerliğini gösteren kanıtlar oluşturmayı hedeflenmiştir.

Orta Doğu Teknik Üniversitesi'nde (ODTÜ) eğitim dili İngilizce'dir. Bu sebeple bu üniversitede okumak isteyen tüm adayların belli bir İngilizce dil seviyesine sahip olmaları istenir. Adaylar yeterli seviyede İngilizce dil becerisine sahip olduklarını kanıtlamak için üniversite yönetimince kabul edilen bir dil yeterlilik sınav sonucunu gösterirler. Bu sınav ODTÜ Yabancı Diller Yüksek Okulu (YDYO) tarafından hazırlanan İngilizce Yeterlilik Sınavı (ODTÜ-İYS) olabileceği gibi, bazı uluslararası kurumların hazırladıkları sınavlar da olabilir; örneğin, TOEFL veya IELTS. Yeterli dil becerisine sahip olmayan adaylar üniversitenin dil okulu olan Temel İngilizce Bölümünde (TİB) bir yıl eğitim aldıktan sonra tekrar İngilizce yeterlilik sınavına girerler.

Adaylar için ODTÜ-İYS hakkında üniversitenin sitesinde sınavın içeriği, yapılış şekli ve puanlaması hakkında bilgi bulunmaktadır. Ayrıca, ODTÜ Kitap Satış Müdürlüğü'nde (Bookstore) *METU-EPE Booklet* adında bir kitapçık bulunmaktadır. Kitapçıkta sınavın biçimi, uygulanma şekli ve örnek sorular vardır. Bunların dışında, sınav notları baz alınarak yapılan bazı istatistik analizlerinin sonuçları dönem dönem TİB'de çalışan öğretim elemanlarına gönderilmektedir. Ayrıca, bazı öğretim elemanlarının zaman zaman ODTÜ-İYS üzerinde küçük çaplı araştırmalar yaptığı bilinmektedir. ODTÜ-İYS'nin geçerliğini inceleyen bir master tezi de bulunmuştur (Ataman, 1999). Ancak, bunları dışında sınav hazırlama komitesi veya yöneticilerin kullanımı için başka bir belge mevcut değildir. Bu durum, sınavın geçerliğine gölge düşürmektedir. Bu yüzden sınavın teorik altyapısının, içeriğinin, ve diğer önemli kriterlerinin sistemli bir araştırma ile incelenmesi gerektiğini göstermektedir.

Yüksek etkili sınavlarda adayların notu üzerinden yapılan çıkarımlar önemli sonuçlara yol açar. ODTÜ-İYS yüksek etkili bir sınav olduğundan sonuç bazlı verilen kararlar da önemlidir. ODTÜ'ye yeni kaydolan öğrenciler ODTÜ-İYS'de geçerli not almaları durumunda kazandıkları programın birinci sınıfına devam etme hakkı kazanırlar; aksi durumunda, bir sene İngilizce eğitim almaları gerekir. Benzer şekilde, ODTÜ'de yüksek lisans yapmak isteyen adayların ODTÜ-İYS'den veya geçerli kabul edilen yabancı sınavlardan birinden (TOEFL veya IELTS) başvurmak istedikleri programın belirlediği aralıkta puan aldıklarını göstermeleri gerekir. Yeterli puanı alamayan adaylar, yüksek lisans başvurusu yapamazlar. Bunlar önemli sonuçlardır, bu yüzden ODTÜ-İYS'nin sonuçları esas alınarak verilen kararların geçerlemesinin yapılması gerekir.

Geçerleme çalışması geçerlilik kavramının işlevselleştirilmesidir. Geçerlik kavramı 1940'larda korelasyon ve faktör analizi üzerine temellenmişti.1955'te Cronbach ve Meehl'in (1955) yaklaşımı geçerlik kavramını radikal bir şekilde değiştirdi. Yeni tanımda geçerlik, sınava ait bir özellik değil sınavı meydana getiren bileşenlerin arasındaki içkin ilişkiyi gösteren birimsel bir kavram olarak ortaya kondu. Bu yeni yaklaşım, sınav soruları (veya görevleri) ile sınavı alanlar arasındaki etkileşim gibi değişik boyutlarının analizini gerektiriyordu. Bu yaklaşım, ayrıca, sınavın geçerliğini araştırmak için de bir çerçeve sundu (Sireci, 2009). 1989'da yayınlanan ikinci bir makale bu konuda yeni ufuklar açtı: bu makalede Messick (1989b) geçerlik savlarının sınavla ilgili değil sınav sonuçlarının değerlendirilmesi ile ilgili olduğunu ortaya attı. Buna göre, bir sınavın geçerliğini savlamak için sınav sonuçları baz alınarak yapılan çıkarımların uygunluğu ve yeterliği için hem kuramsal hem de görgül kanıtlara gereksinim vardır.

Messick'in (1989) geçerlik kavramına yaklaşımı temel alındığında, ODTÜ-İYS yüksek etkili bir sınav olduğundan sınav notlarının ölçülen becerilen geçerli ve güvenilir bir göstergesi olduğuna dair kanıt toplamak ve bunları sunmak sınav geliştiricilerinin sorumluluğundadır. ODTÜ-İYS'nin kuramsal altyapısı veya sınav notlarının değerlendirme kriterleri bilinmemektedir. Ayrıca sınavın hedeflerini veya sınav yazarlarının / geliştiricilerinin uyguladıkları ilkeleri belirleyen bir belge bulunamamıştır. ODTÜ-İYS'nin sonuçları sadece sınavı alan adayları değil, TİB ve Modern Diller Bölümü (MDB) gibi İngilizce eğitimi veren bölümleri, fakülteleri, ve diğer öğretim elemanlarını da etkilemektedir. Bu yüzden, ODTÜ-İYS'nin geçerlemesini yapmak üzere sistemli araştırma gerektiği ortaya çıkmıştır. Bu amaçla, Weir'in (2005) sosyal-bilişsel geçerleme çerçevesi kullanılarak ODTÜ-İYS'nin okuma bölümünün bağlamsal, bilişsel ve notlama geçerliği araştırılmıştır.

Bu araştırmada kullanılan sosyal-bilişsel geçerleme çerçevesi sınav geliştirmede uygulanması gereken adımları zamansal bir çerçeve içinde sunmaktadır. Bun göre, sınav geliştirme adımları sınav öncesi işlemler ve sınav sonrası işlemler olarak ikiye ayrılmıştır. Bu araştırmanın birinci ve ikinci soruları sınav öncesi işlemleri irdelerken, üçüncü soru da sınav sonrası notlama geçerliğini sorgulamıştır. Buna göre, araştırma soruları şunlardır:

1. Akademik okuma becerisinin bir sınav kurgusu olarak kavramsal ve işlevsel tanımı nedir?
2. Okuma sınav kurgusunun temelindeki bilişsel süreçler a) geçmişe dönük anımsama ve b) sınav anında içebakış yöntemleri ile nelerdir?
3. Sınav madde parametreleri ne ölçüde sınavın geçerlik savlarını desteklemektedir?

Bu çalışma sınav geliştirmede daha sistemli bir yaklaşım için örnek olacaktır. Ölçme bir çalışma alanı olarak henüz çok popüler değildir çünkü üniversitelerde ölçme üzerine verilen ders sayıları ve ders içerikleri yetersizdir (Hatipoğlu, 2015). Sınav geliştirme konusunda sınav yazarları genellikle deneyimli öğretim elemanları arasından seçilir. Öğretme deneyimi sınav geliştirmede önemli bir kaynak olmasına rağmen yeterli değildir. Ölçme okur yazarlığında, sınav geliştirme ve geçerleme çalışması bilgisi içerik bilgisi kadar önemlidir. Bu anlamda, bu çalışma dil öğretimi alanında ölçmenin önemini ortaya koyacak ve sınav geliştirmede değişik yaklaşım işlemler hakkında farkındalık yaratacaktır.

Küresel anlamda bu çalışma geçerli mi amacıyla bir çerçeve kullanımı konusunda var olan bilgi birikimine katkı sağlayacaktır. Bir geçerleme / sınav geliştirme çerçevesi kullanmak bir ölçüm aracının geliştirilmesinde sağlam bir temel sunar. Ayrıca bir çerçevenin farklı bağlamlarda kullanılması bu çerçevenin tüm yönleriyle başka ortamlarda da geçerli olup olmadığını göstermiştir. Bu sebeple, bu çalışma İngilizcenin anadil olduğunu bir ortamda geliştirilen bir sınav geliştirme ve geçerleme aracının, İngilizcenin yabancı dil olduğu bir ortamda kullanılmasıyla okuma becerisinin kavramsal olarak tanımlanmasına katkıda bulunmuştur.

Kaynak taraması

Modern anlamda geçerlik kavramı ilk olarak psikometri yazınında 1920 lerde ortaya çıktı ve temel olarak herhangi bir korelasyonu geçerlilik göstergesi olarak kabul eden pragmatik bir yaklaşıma sahipti (Sireci, 2009). 20. yüzyılın ortalarına doğru test geçerlemesinin sadece istatistiksel yöntemlerle yapılmasından duyulan rahatsızlık başka bakış açıları ortaya çıkardı. Rulon (1946) bir sınavın geçerli olup olmadığını söyleyebilmek için önce sınavın amacını belirlenmesi ve geçerleme kanıtlarının tanımlanması gerektiğini söyledi. Kısa bir zaman sonra, Amerikan Psikoloji Birliği (APA) sınav standartlarına tanımlamak için bir komite kurdu. Bu komite geçerliği üç kategoride tanımladı: içerik, ölçütsel ve yapısal geçerlik. Bu komiteden iki üye (Cronbach ve Meehl) bir kaç yıl sonra yapısal geçerlik kavramını geliştirdiler (1955). Onlara göre, yapısal geçerlik kavramı işlevsel olarak tanımlanamayan niteliklerin ölçülmesinde kullanılmalıydı. Ancak, yapısal geçerliğin eğitim alanındaki tüm sınavlarda uygulanabileceği kabul edildi (Loevinger, 1957; Messick, 1989b).

Geçerlik teorisine önemli katkılar yapmış olan Messick (1989) içerik, kriter ve sonuç faktörlerini barındıran bir kapsamlı bir çerçeve tanımladı ve bir geçerlik tanımı önerdi. Bu tanıma göre geçerlik görgül kanıtların ve kuramsal gerekçelerin sınav notları üzerinden yapılan çıkarımların yeterlilik ve uygunluğunu ne derecede desteklediğini gösteren bütüncül değerlendirme yargısıdır (p.13). Bu görüşte, geçerlik sınavın bir özelliği değil sınav sonuçlarının doğrulanabilme derecesi olarak ele alınıyor.

Messick'in (1989b) geçerleme tanımı sınav geliştirme işlemlerini büyük ölçüde etkiledi (Chapelle, 1999). Ancak Messick'in (1989b) oldukça karmaşık bir kavramsal tanıma sahip geçerleme çerçevesi daha sonra pratik ve daha kolay uygulanabilir çerçevelerin ortaya çıkmasına yol açtı. Bu çerçeveleden biri de Weir (2005) tarafından tanımlanan sosyal-bilişsel çerçevedir. Bu çerçeve dört beceri için (okuma, dinleme, yazma, ve konuşma) kriter bazlı bir model sunar. Her bir beceri modelinde beş alan tanımlanmıştır: bağlam, bilişsel, notlama, sonuçsal ve kriter bazlı geçerlik. Bu çerçevede sınav öncesi ve sınav sonrası olmak üzere yapılması gereken işlemler ikiye ayrılmıştır. Bağlam ve bilişsel geçerlilik sınav öncesi işlemlerle, notlama, sonuçsal ve kriter bazlı geçerlik sınav sonrası yapılan işlemlerle belirlenmektedir.

Bu çalışmada, sınav öncesi bağlam ve bilişsel geçerlik ve sınav sonrası notlama geçerliği araştırılmıştır. Weir (2005) bu üç geçerlik kavramının, geçerlik teorisinin tepe noktasında yer alan yapısal geçerliği oluşturduğunu iddia eder.

YÖNTEM

Bu çalışma tamamlayıcı, karma yöntemli bir çalışmadır. Bu çalışmada üç araştırma sorusu vardır. Birinci araştırma sorusu, "Akademik okuma becerisi kavramsal ve işlevsel olarak nasıl tanımlanır?" kavramsal bir sorudur. Bu soruya cevap vermek için kaynak taraması yapılmış ve 2013-2015 yıllarında YDYO bünyesinde yapılmış olan ihtiyaç analizi çalışmasının sonuçları incelenerek okuma becerisinin parametreleri belirlenmiş ve sınav tanımlamaları dosyasına yazılmıştır. İkinci araştırma sorusu, "Okuma sınav kurgusunun temelindeki bilişsel süreçler a) geçmişe dönük anımsama ve b) sınav anında içebakış yöntemleri ile nelerdir?" keşifçi bir sorudur ve geçmişe dönük ve içe bakış yöntemleri ile sınavı alanların bilişsel süreçlerini incelemeyi hedeflemiştir. Üçüncü araştırma sorusu, "Sınav madde parametreleri ne ölçüde sınavın geçerlik savlarını desteklemektedir?", yine keşifçi bir sorudur ve okuma sınavını alanların her bir soruya verdikleri cevabı ve sınavdan aldıkları notları istatistiksel olarak incelemeyi hedeflemiştir.

Çalışma üç aşamadan oluşmuştur. Birinci aşamada birinci araştırma sorusunun cevabını vermek üzere akademik okuma üzerine kaynak taraması ve ihtiyaç analizi incelemesi yapılmış ve ardından sınav komitesi ile birlikte yeni hazırlanacak okuma sınavında kullanılacak kriterler belirlenmiştir. Bu kriterler daha sonra sınav tanımlamaları dosyasına yazılmıştır. Bu sınav tanımlaması sınav komitesi tarafından yeni okuma sınavının hazırlanmasında kullanılmıştır. Çalışmanın ikinci aşamasında, hazırlanmış olan okuma sınavı birinci sürümünün dört ayrı bölümü TİB öğrencisi olan 100'er katılımcıya birer küçük test olarak uygulanmış ve uygulama aşamasında katılımcıların geriye dönük inceleme protokol formu doldurmaları istenmiştir. Toplanan veri, sınav sonuçları ve geriye dönük inceleme protokol formu nicelik olarak incelenmiştir. Elde edilen sonuçlar sınav komitesi ile tartışılmış, ve gerekli değişikliklere karar verdikten sonra okuma sınavı ikinci sürümü üretilmiştir. Çalışmanın üçüncü ve son aşamasında, okuma sınavı ikinci sürümü, gene TİB'den 26 katılımcıya bir bütün olarak uygulanmış, ve uygulama sırasında sesli düşünme tekniği uygulamaları istenmiştir. Katılımcıların düşünceleri kaydedilmiş ve daha sonra yazıya çevrilmiştir. Sınav sorularına verilen cevaplar ve sınav notları üçüncü araştırma sorusuna cevap vermek üzere sayısal olarak incelenirken, sesli düşünme tekniği ile elde edilen veri hem nicelik hem de nitelik olarak incelenmiştir. Sonuçlar okuma sınavı ikinci sürümünde yapılması gereken değişikliklere ışık tutmuş ve her bir aşamada elde edilen veriler okuma sınavının bağlam, bilişsel ve notlama geçerliliği için kanıt sağlamıştır.

Örnekleme ve kullanılan araçlar

Bu çalışmanın her iki aşamasının katılımcıları TİB öğrencileri arasından seçilmiştir. Okuma sınavının birinci sürümünün her bir görevi (alt sınavları) beş ayrı gruptan (Pre-intermediate, Intermediate, Upper-intermediate, Advanced ve Repeat) yaklaşık 100'er öğrenciye uygulanmıştır. Örnekleme yöntemi kademeli gelişigüzel örneklemedir. Bu aşamadaki toplam katılımcı sayısı 400'dür. Bu aşamada kullanılan araçlar, Okuma Sınavı V1 (birinci versiyon) ve geriye dönük inceleme formudur. Okuma Sınavı V1 sınav komitesi tarafından hazırlanmıştır. Geriye dönük inceleme formu Weir, Hawkey, Green ve Devi (2009) çalışmasından uyarlanmıştır.

Çalışmanın üçüncü aşamasında verilen Okuma Sınavı V2 yine TİB öğrencilerine uygulanmıştır. Bu sürüm, seçilen metodoloji sebebiyle - sesli düşünme - 26 katılımcıya uygulanmıştır. Katılımcılar kota örnekleme yöntemiyle seçilmiştir. Bu aşamada kullanılan araçlar, Okuma Sınavı V2 ve içe dönük inceleme formudur. Okuma Sınavı V2 sınav komitesi tarafından hazırlanmıştır. İçe dönük inceleme stratejileri Cohen ve Upton (2006) çalışmasından uyarlanmıştır.

Veri Analizi

Okuma Sınavı V1 sayısal veri sağladığından sınav sorularına verilen cevaplar ve sınav notları üzerinde klasik sınav teorisi analizi uygulanmıştır. Okuma sınavı sırasında doldurulan geriye dönük inceleme formu da sayısal veri sağlamıştır. Bu verilere betimsel ve çıkarımsal istatistik analizleri Microsoft Excel (2016) ve IBM SPSS (V24) kullanılarak yapılmıştır.

Okuma Sınavı V2 sayısal veri sağlamış, ve sınav sorularına verilen cevaplar ve sınav notları üzerinde klasik sınav teorisi analizi uygulanmıştır. Okuma sınavı sırasında sesli düşünme tekniği ile sözel veri toplanmıştır. Bu veri önce yazıya dönüştürülmüş, daha sonra bilgisayar ortamına aktarılarak MaxQDA (V.16) programında Cohen ve Upton'ın (2006) içe dönük inceleme stratejileri, ve benim eklediğim 11 yeni strateji ile kodlanarak kategorize edilmiştir. Esasta niteliksel veri olmasına rağmen kullanılan stratejilerin sıklık derecelerinin anlaşılmasının çalışmayı anlamlandıracağı düşüncesi ile bu veri de nicelik olarak incelenmiş ve sonuçlar sayısal olarak verilmiştir.

BULGULAR

Araştırma sorusu 1: Akademik okuma becerisinin bir sınav kurgusu olarak kavramsal ve işlevsel tanımı nedir?

Kaynak yazın tarama sonuçları

Kaynak yazın okuma becerisinin tanımlanmasında pek çok değişik yaklaşımın olduğunu göstermektedir. Bu yaklaşımlar temel olarak iki grup altında toplanabilir: Süreç ve bileşen modelleri. Süreç modelleri aşağıdan yukarı, yukarıdan aşağı ve etkileşimli olmak üzere üç tanedir. Aşağıdan yukarı süreç modelinde okuma önce harflerin, sonra hecelerin ve kelimelerin deşifre edilmesiyle üst kademelere doğru anlamlandırmaya başlayarak okuma sürecine verilen isimdir. Düşük yeterlik seviyesine sahip okuyucuların daha çok aşağıdan yukarı süreç modelini kullandığı savlanmıştır (Treiman, 2017). Yukarıdan aşağı süreçte okuma tüm metinden başlayarak aşağı doğru iner. Bağlam ve konuya özel bilgi metnin anlaşılmasına ve anlamlandırılmasına yardımcı olur. Bu gruptaki sonunda model etkileşimli modeldir. Yapılan araştırmalar aşağıdan yukarı ve yukarıdan aşağı modellerin okuma sürecini tam olarak açıklayamadığını ortaya koymuş ve bu iki model arasında bir etkileşim olduğunu göstermiştir. Günümüzde okumanın hem aşağıdan yukarı görsel bilginin işlenmesi hem de dünya bilgisi kullanarak yukarıdan aşağıya metnin anlamlandırılması işlemlerinin aynı ayna yürütüldüğü düşünülmektedir (Faerch ve Kasper, 1986).

Bileşen modelleri okumanın bileşenlere ayrılarak öğretilebileceğine ve sınanabileceğini öngörür. Okuma becerisini alt süreçlerine ayırmak sınav performansının faktör analizine tabi tutulmasına dayanır (Johnston, 1981). Bu yöntem, farklı okuma görevleri veya sınav maddelerinin okumayı oluşturan bir kaç faktöre yüklenmesi ile açıklanır. Kaynak yazında okuma becerisinin bölünebileceği görüşü çeşitli araştırmalarala hem desteklenmiş hem de çürütülmüştür.

Bu çalışmada okuma bilişsel bir süreç olarak kabul edilmiştir. Bunun için, okumayı tanımlarken Khalifa ve Weir (2009) tarafından önerilen okumada bilişsel süreç modeli kullanılmıştır. Bu modelde amaç belirleyici olan üstbilişsel etkinlik bölümü, bu bölümde belirlenen amaçlara uygun olarak etkinleşen bilişsel süreçlerin olduğu ana işlem merkezi ve gereksinim duyulduğunda kullanılan bilgi merkezi bulunmaktadır. Üstbilişsel etkinlik alanındaki amaç belirleyicinin altında okuma tipleri iki boyutlu olarak belirlenmiştir. Bunlar dikkatli - hızlı okuma ve yerel - küresel okuma olarak adlandırılmıştır. Urquhart ve Weir'in (1998) tanımladığı bu iki boyutlu okuma tipi tanımı okuma sınavının geliştirilmesinde baz alınmıştır. Buna göre yerel dikkatli okuma cümle bazında okuduğunu anlamayı, küresel dikkatli okuma metin bazında okuduğunu anlamayı, yerel hızlı okuma belirli bir kelime, özel isim, sayı veya tarihi bulmak için tarama yapmayı ve küresel hızlı okuma metnin konusunu ve belki de ana fikirlerini genel olarak anlamak için metne hızlıca göz gezdirmeyi tanımlar.

İhtiyaç analizi sonuçları

İhtiyaç analizi çalışması 2013 - 2015 yılları arasında ODTÜ'deki beş fakültenin öğretim elemaları, TİB ve MDB'den öğretim elemanları, TİB öğrencileri ve fakültelerdeki öğrencilerden alınan nitel ve nicel verilerin incelenmesi ile yapılmıştır. Bunların yanısıra okuma kriterlerinin belirlenebilmesi için bölümlerde okutulan kitaplar, verilen sınavların örnekleri, ders izlenceleri ve öğrencilere verilen ödevler incelenerek okuma kriterleri oluşturulmuştur. Buna göre, okuma becerisi üç ana başlık altında açıklanmıştır.

Bilgi ve düşünce için okuma: Bu okumada amaç açık veya örtülü olarak verilen ana fikirleri ve diğer temel bilgiyi anlamaktır. Okuyucu ayrıca metindeki düşünceleri de ayırt etmelidir. Bu tür okuma bir sınav veya ödev için öğrenmek amacıyla yapılan okumadır. Bu okumada kullanılan metinler okul kitapları, diğer kitaplar, makaleler, hocaların notları, sunumlar vb. olabilir.

Oryantasyon için okuma: Bu okumada amaç önceden belirlenmiş bir konuyu uzun ve karmaşık bir metnin içinde bulma ve ilgili bölümleri dikkatli okuyarak anlamaktır. Kullanılan metinler kitaplar, makaleler, internet ortamında bulunabilecek alana özgü metinler ve teknik raporlar olabilir.

Yönerge okuma: Burada amaç okunan her bir cümlenin detaylı anlaşılmasıdır. Kullanılan metinler bir cümleden bir kaç cümleye çıkabilen sınav soruları, sınav yönergeleri, ve sınıf için uygulamalarda kullanılan diğer yönergeler (örneğin, rapor hazırlama yönergesi) olabilir.

Yazın taraması ve gereksinim analizi incelemesi sonucunda elde edilen bilgilerin Urquhart ve Weir'in (1998) okuma modeline uyarlanması ile ortaya çıkan okuma sınavında ağırlıklı olarak küresel dikkatli okumanın, ayrıca hızlı küresel ve hızlı yerel okumanın sınanması gerektiği ortaya çıkmıştır. Weir'in (2005) sosyal-bilişsel geçerleme çerçevesi sınav öncesi işlemlerde yer alan bağlam geçerliği için belirlenmesi gereken kriterleri vermektedir. Hazırlanan okuma sınavında yazın taraması ve gereksinim analizi incelemesi sonucunda belirlenen kriterler sınav tanımlama dosyasına yazılarak sınavın birinci versiyonu hazırlanmıştır.

Araştırma sorusu 2: Okuma sınav kurgusunun temelindeki bilişsel süreçler a) geçmişe dönük anımsama ve b) sınav anında içebakış yöntemleri ile nelerdir?

İkinci araştırma sorusunun birinci alt sorusu geçmişe dönük anımsama ile okuma sınavının etkinleştirdiği bilişsel süreçleri açığa çıkarmayı hedeflemiştir. Bu amaçla, uygulanan Okuma Sınavı V1 ve geçmişe dönük anımsama yöntemi ile doldurulan protokol formu analiz edilmiş ve aşağıdaki sonuçlar bulunmuştur.

Protokol formunun B bölümü katılımcılara sınava başladıklarında sorulara bakmadan önce metin üzerinde bir önizleme yapıp yapmadıklarını soruyordu. Formlardaki bilgiye göre alt sınavlardan alınan bilgilenin bir örüntü göstermediği, katılımcıların ortalamada her bir seçeceği (*yavaş ve dikkatli önizleme, hızlı ve seçerek önizleme,* ve *önizleme yok*) birbirine yakın oranlarda seçtikleri görülmüştür.

Protokol formunun C bölümü, katılımcıların her bir sınav sorusunu cevaplarken kullandıkları sınav cevaplama stratejilerini anlamaya yönelikti. Bu bölümden elde edilen sonuçlara göre, alt sınavların üç tanesinde en sık kullanılan dört strateji aynıydı. Bu stratejiler, *yavaş ve dikkatli okuma*, *bir bölümü tekrar dikkatli okuma*, *kelimeyi arama ve eşleştirme* ve *kelime bilgisi kullanma* stratejileriydi. Bir alt sınavda ek olarak *metnin anahtar bölümlerini okuma* stratejisi ilk dört strateji arasına girmiştir.

Protokol formunun D bölümü katılımcıların doğru olduğunu düşündükleri cevabı nerede bulduklarını sorguluyordu. Bu bölümün analizi, her dört alt sınavda katılımcıların doğru olduğunu düşündükleri cevabı iki veya daha fazla cümledeki anlamdan çıkardıklarını gösterdi. İkinci sırada, cevabın tüm metinden çıktığı, üçüncü sırada ise tek bir cümleden cevabın bulunduğu söylendi.

Özet olarak geçmişe dönük anımsama protokol formu katılımcıların sınav sorularına cevap ararken en sık yavaş ve dikkatli okuma stratejileri kullandığını ve sınav sorularının cevaplarını küresel okuma yaparak bulduklarını gösterdi.

İkinci araştırma sorusunun ikinci alt sorusu içe bakış yöntemi ile okuma sınavının etkinleştirdiği bilişsel süreçleri açığa çıkarmayı hedeflemiştir. Bunu yaparken her bir soru tipinin gerektirdiği stratejiler ayrı incelenmiştir. Buna göre sonuçlar eşleştirme, kelime anlamı, genel düzeyde anlama, arayarak okuma ve hızlı okuma başlıkları altında ayrı ayrı sunulmuştur.

Tablo 1 Tüm soru tipleri için stratejilerin kullanım sıklıkları:

		Dikkatli Yerel Okuma (Çoktan Seçmeli) Kelime	Dikkatli Geniş Çaplı Okuma (Çoktan Seçmeli) Genel Anlama	Arayarak Okuma (Kısa Cevap)	Hızlı Okuma (Eşleştirme)
	Okuma Stratejileri[6]				
RS1	Plan a goal	0,50	0,57	0,13	2,00
RS2	Make a mental note	0,50	0,71	0,13	0,50
RS4	Read text carefully	-	0,14	-	0,50
RS6	Read a portion carefully	10,50	16,50	19,38	12,33
RS7	Scan	2,00	3,21	19,38	2,33
RS8	Look for markers of meaning	0,50	0,64	1,63	0,33
RS9	Repeat, paraphrase	2,50	3,21	0,88	3,33
RS10	Identify unknown word	2,00	2,14	0,25	2,67
RS11	Identify unknown sentence	0,50	0,07	0,25	0,33
RS12	Reread	1,50	0,86	2,13	1,00
RS13	Ask overall meaning	2,50	1,36	0,75	1,67
RS14	Monitor understanding	0,50	1,64	0,88	1,83
RS15	Adjust comprehension (previous)	-	1,29	0,88	1,67
RS16	Adjust comprehension (new)	0,50	2,64	1,63	3,83
RS17	Confirm understanding	1,50	2,57	1,38	1,67
RS19	Identify keyword	-	0,86	0,50	2,33
RS20	Search main idea	0,50	0,50	0,88	1,50
RS21	Use discourse knowledge	-	0,57	0,63	0,67
RS22	Use organization knowledge	-	0,71	0,75	0,33
RS23	Use logical connectors	-	0,14	0,38	0,17
RS24	Read ahead	-	1,00	0,75	0,33
RS25	Go back	3,50	1,50	0,90	0,17
RS26	Verify referent	0,50	0,36	0,13	-
RS27	Infer meaning (internal)	-	0,50	0,13	0,17
RS28	Infer meaning (external)	3,00	0,14	0,13	0,67
RSNEW1	Skim	1,50	1,14	2,75	2,00
RSNEW2	Read text again	9,00	7,43	4,38	3,67

[6] Bu okuma stratejileri Cohen ve Upton (2006) çalışmasından uyarlanmış ve yazıldığı dilde (İngilizce) kullanılmıştır. Stratejilerin çevirisinin ancak geçerliği yapıldıktan sonra anlamlı olacağı bilindiğinden yazıldığı dilde bırakılmıştır.

Tablo 1'in devamı:

	Soru Cevaplama Stratejileri				
TM1	Reread question	1,00	5,64	19,38	0,50
TM2	Paraphrase question	-	0,43	2,38	-
TM3	Wrestle with question intent	-	0,36	1,75	0,33
TM4	Read the question & options	7,50	9,71	0,13	2,50
TM5	Read the question & text	11,00	11,93	17,38	1,67
TM6	Predict own answer after reading	1,00	0,43	0,13	0,33
TM7	Predict own answer before reading	0,50	0,43	-	0,17
TM9	Identify unknown vocabulary	0,50	0,14	-	-
TM11	Consider a familiar option	-	0,29	-	0,83
TM12	Select option though uncertain	3,00	1,14	-	1,83
TM15	Drag the option to the sentence	1,50	1,07	-	-
TM17	Wrestle with option meaning	0,50	0,79	-	0,17
TM18	Make a guess	0,50	0,71	0,25	0,17
TM20	Locate vocabulary in context	5,50	0,21	-	-
TM22	Select option (meaning)	6,50	6,00	0,13	4,17
TM24	Select option (meaning/elimination)	3,50	6,64	-	0,90
TM25	Select option (elimination)	-	0,50	-	-
TM26	Select option (discourse)	-	1,00	0,13	-
TMNEW1	Identify answer (keyword)	-	-	0,88	-
TMNEW2	Identify answer (meaning)	-	-	5,13	-
TMNEW3	Identify section (content)	-	-	3,13	-
TMNEW4	Identify section (keyword)	-	-	2,75	-
TMNEW5	Identify section (discourse)	-	-	2,00	-
TMNEW6	Identify section (subtitles)	-	-	3,13	-
TMNEW7	Identify keywords in Q	0,50	0,93	8,25	1,17
TMNEW8	Identify unknown vocabulary	1,50	0,71	1,50	0,33
	Deneyim Stratejileri				
TW1	Elimination	0,50	0,57	-	-
TW3	Select by keyword	-	0,29	-	1,00
TWNEW1	Use item sequence information	-	-	0,75	-

Bu sonuçlara göre kelime sorularını cevaplarken en sık kullanılan ilk üç okuma stratejisi şunlardır: *metni dikkatle okumak* (RS6), *metnin bir kısmını tekrar dikkatle okumak (*RSNEW2), ve *okuduğunu anlayabilmek için metnin diğer bölümlerinden yararlanmak* (RS25).

Genel anlama sorularına cevap verirken en sık kullanılan ilk üç okuma stratejisi şunlardır: *metni dikkatle okumak* (RS6), *metnin bir kısmını tekrar dikkatle okumak* (RSNEW2), ve *bir bilgiyi bulmak için metni taramak* (RS7).

Başlık eşleme sorularını cevaplarken en sık kullanılan ilk üç okuma stratejisi şunlardır: *metni dikkatle okumak* (RS6), *yeni okunan bilginin daha önceden okuduklarını destekleyip desteklemediğini anlamak* (RS16) ve *metnin bir kısmını tekrar dikkatle okumak* (RSNEW2).

Arayarak okuma sorularını cevaplarken en sık kullanılan ilk üç okuma stratejisi şunlardır: *metni dikkatle okumak* (RS6), *bir bilgiyi bulmak için metni taramak* (RS7) ve *metnin bir kısmını tekrar dikkatle okumak* (RSNEW2).

Soru cevaplama stratejilerinde ise kelime sorularına cevap verirken en sık kullanılan ilk üç soru cevaplama stratejisi şunlardır: *soruyu okuduktan sonra seçenekleri gözden geçirirken metinde cevabı aramak* (TM5), *soruyu okuduktan sonra önce seçeneklere göz gezdirmek* (TM4) ve bir seçeneği *kelime, cümle, paragraf veya metnin genel anlamından dolayı seçmek*.

Genel anlama sorularına cevap verirken en sık kullanılan ilk üç soru cevaplama stratejisi şunlardır: *soruyu okuduktan sonra seçenekleri gözden geçirirken metinde cevabı aramak* (TM5), *soruyu okuduktan sonra önce seçeneklere göz gezdirmek* (TM4) ve *metnin anlamına uymayan seçenekleri eleyerek cevabı seçmek* (TM24).

Arayarak okuma sorularına cevap verirken en sık kullanılan ilk üç soru cevaplama stratejisi şunlardır: *daha iyi anlayabilmek için soruyu tekrar okumak* (TM1), *soruyu okuduktan sonra seçenekleri gözden geçirirken metinde cevabı aramak* (TM5), ve *soruda anahtar kelime bulmak* (TMNEW7).

Başlık eşleştirme sorularına cevap verirken en sık kullanılan ilk üç soru cevaplama stratejisi şunlardır: *kelime, cümle, paragraf veya metindeki genel anlama göre cevabı seçmek* (TM22), *soruyu okuduktan sonra önce seçeneklere göz gezdirmek* (TM4), ve *soruyu okuduktan sonra seçenekleri gözden geçirirken metinde cevabı aramak* (TM5).

Bu sonuçlara göre, her dört tip okuma sorusu en çok geniş çaplı dikkatli okuma, ardından belli bir bilgiye ulaşmak için metni tarama stratejilerinin kullanımına yol açmıştır.

Okuma Sınavı V2'den yüksek ve düşük not alan katılımcıların kullandıkları stratejiler karşılaştırıldığında ise şu sonuçlar bulunmuştur: Yüksek not alan katılımcıların kullandığı okuma ve soru cevaplama stratejileri soru tiplerine daha uygundu. Eşleştirme sorularına cevap verirken, bu katılımcılar bazı stratejileri (RS6, RS16, TM5, TM22, RS9, RS19, RSNEW2, RS13 VE RS10) benzer oranlarda kullandılar (kullanım sıklığı 0.33 – 1.67 arası). Oysa düşük not alan katılımcılar esas olarak dikkatli okuma (RS6 kullanım sıklığı=4.83) stratejisini kullanırken, diğerlerini daha düşük oranda kullandılar (kullanım sıklığı 1.17 – 0 arası). Eşleştirme sorularında hızlı okuma tekniklerinin kullanımı beklendiğinden, düşük not alan katılımcıların gereken stratejileri seçmede başarılı olamadıkları söylenebilir.

Genel anlama sorularında odak noktası tüm metnin detaylı bir şekilde okunup anlaşılması üzerineydi. Buna uygun olarak tüm katılımcılar dikkatli okuma stratejisi kullandılar. Ancak düşük not alan katılımcılar metni okurken daha kısa zaman aralıkları ile durup, soruyu tekrar okudular.

Kelime sorularında katılımcıların bağlamdaki ipuçlarını kullanarak sorulan kelimenin anlamaları bekleniyordu. Bu yüzden kısa bir veya iki cümle okumalarının yeterli olacağı düşünülmüştü. Yüksek not alan katılımcıların pek çoğu bu şekilde doğru cevaba ulaşırken, düşük not alan katılımcılar soru cevaplama stratejisi kullanarak doğru olduğunu düşündükleri cevabı seçtiler.

Arayarak okumada yüksek not alan katılımcılar metindeki başlık ve alt başlıkları kullanarak okumaları gereken bölgeyi seçmede başarılı oldular. Düşük not alan katılımcılar daha çok kelime tarama stratejisi kullanarak okumaları gereken bölümü bulmaya çalıştılar. Kelime bazında taramada başarısız olduklarında dikkatli okuma stratejisi uyguladılar.

Araştırma sorusu 3: Sınav madde parametreleri ne ölçüde sınavın geçerlik savlarını desteklemektedir?

Okuma Sınavı V1sonuçları

Okuma Sınavı V1 katılımcılara dört ayrı sınav görevi olarak uygulandığından sonuçlar da ayrı ayrı verilmiştir. Buna göre sınavın betimleyici istatistik değerleri şöyledir:

Tablo2 Betimleyici istatistik analizleri

	Alt sınav 1	**Alt sınav 2**	**Alt sınav 3**	**Alt sınav 4**
Ortalama	3.0 (38%)	3.3 (47%)	5.9 (74%)	4.0 (57%)
Standart Sapma	1.61	1.43	1.57	1.58
Varyans	2.6	2.0	2.5	2.50
Çarpıklık	0.50	0.03	-1.19	-0.06
Basıklık	-0.19	-0.16	1.97	-0.76
Alpha Katsayısı	0.43	0.25	0.51	0.38
Aralık	7	7	8	6
En az	0	0	0	1
En çok	8	7	8	7

Bu sonuçlara göre, en zor alt sınav 1, en kolay alt sınav 3 olmuştur. Alt sınav 1'deki 8 sorunun 6 tanesi eşleştirme, 2 tanesi genel anlama sorusu idi. Alt sınav 3'te 8 sorunun 6 tanesi Evet/Hayır sorusu, diğer 2 tanesi genel anlama sorusu idi. Okuma Sınavı V1'in Klasik Sınav Teorisi'ne göre madde analizleri aşağıda verilmiştir.

Tablo 3 Okuma Sınavı V1madde analizleri

Madde #	Madde Tipi	Madde Kolaylığı (IF)	Ayırıcılık Göstergesi (*d*)	Nokta İki Serili Korrelasyon (r_{pbi})
Alt sınav 1				
1	Eşleştirme	0.82	0.41	0.36
2	Eşleştirme	0.37	0.70	0.61
3	Eşleştirme	0.28	0.44	0.53
4	Eşleştirme	0.14	0.26	0.38
5	Eşleştirme	0.27	0.41	0.53
6	Eşleştirme	0.32	0.56	0.50
7	Çoktan Seç.	0.44	0.56	0.34
8	Çoktan Seç.	0.43	0.44	0.33
Alt sınav 2				
1	Çoktan Seç.	0.33	0.56	0.39
2	Çoktan Seç.	0.47	0.52	0.37
3	Çoktan Seç.	0.33	0.26	0.36
4	Çoktan Seç.	0.75	0.30	0.42
5	Çoktan Seç.	0.36	0.33	0.35
6	Çoktan Seç.	0.43	0.15	0.28
7	Çoktan Seç.	0.57	0.07	0.29
Alt sınav 3				
1	Evet/Hayır	0.89	0.17	0.27
2	Evet/Hayır	0.76	0.62	0.55
3	Evet/Hayır	0.85	0.45	0.42
4	Evet/Hayır	0.87	0.38	0.43
5	Evet/Hayır	0.59	0.31	0.34
6	Evet/Hayır	0.92	0.21	0.27
7	Çoktan Seç.	0.56	0.90	0.48
8	Çoktan Seç.	0.85	0.24	0.31
Alt sınav 4				
1	Çoktan Seç.	0.37	0.56	0.18
2	Çoktan Seç.	0.51	0.70	0.37
3	Çoktan Seç.	0.66	0.44	0.16
4	Çoktan Seç.	0.65	0.52	0.32
5	Çoktan Seç.	0.58	0.48	0.32
6	Çoktan Seç.	0.61	0.67	0.31
7	Çoktan Seç.	0.59	0.48	0.26

Tabloda görüldüğü üzere, Alt Sınav 1'de eşleştirme sorularının tamamı belirlenmiş olan IF aralığı (0.40 – 0.80) dışında kalmıştır. Ayırıcılık göstergeleri ve korelasyon değerleri yeterli bulunmuştur. Alt Sınav 2'de 3 çoktan seçmeli soru beklenenden daha zor çıkmış, ayrıca ayırıcılık göstergeleri ve nokta iki serili korelasyon değerleri de 2 soru için yetersiz kalmıştır. Alt Sınav 3'te Evet/Hayır sorularından 4 tanesi beklenen değerlerden daha kolay olmuştur. Ayırıcılık göstergesinde 1 soru, nokta iki serili korelasyonda 2 soru yetersiz parametreler göstermiştir. Alt Sınav 4'te yalnız bir soru beklenen değerden (0.40) düşük kalmış, diğerleri madde kolaylığı açısından yeterli bulunmuştur. Ayırıcılık göstergesi de tüm sorular için yeterlidir. Nokta iki serili korelasyonda 3 sorunun göstergeleri yetersiz bulunmuştur.

Okuma Sınavı V2 sonuçları

Tablo 4 Okuma Sınavı V2 betimleyici istatistik analizleri

	n=26
Ortalama	19.8 66%
Standart Sapma	4.70
Varyans	22.10
Çarpıklık	-0.045
Basıklık	-0.716
Aralık	11 - 28
En az	11.00
En çok	28.00

Bu sonuçlara göre, Okuma Sınavı V2'de katılımcılar önceki versiyona göre daha fazla sayıda soruya doğru cevap vermişlerdir. Not dağılımı normaldir.

Tablo 5 Okuma Sınavı V2'nin Klasik Sınav Teorisine göre madde analizleri

Madde #	Madde Tipi	Madde Kolaylığı (IF)	Ayırıcılık Göstergesi (*d*)	Nokta İki Serili Korrelasyon (r_{pbi})
1	Eşleştirme	0.73	0.50	0.45
2	Eşleştirme	0.96	0.13	0.20
3	Eşleştirme	0.73	0.63	0.54
4	Eşleştirme	0.65	0.50	0.34
5	Eşleştirme	0.88	0.13	0.11
6	Eşleştirme	0.77	0.25	0.28
7	Çoktan Seç.	0.92	0.25	0.42
8	Çoktan Seç.	0.31	0.50	0.32
9	Çoktan Seç.	0.58	0.13	0.29
10	Çoktan Seç.	0.73	0.38	0.38
11	Çoktan Seç.	0.35	0.13	0.12
12	Çoktan Seç.	0.81	0.13	0.02
13	Çoktan Seç.	0.85	-	0.09
14	Çoktan Seç.	0.62	0.13	0.15
15	Çoktan Seç.	0.62	0.13	0.21
16	Çoktan Seç.	0.50	0.38	0.34
17	Çoktan Seç.	0.65	0.38	0.22
18	Çoktan Seç.	0.69	0.63	0.50
19	Çoktan Seç.	0.96	0.13	0.37
20	Çoktan Seç.	0.77	-	(0.12)
21	Çoktan Seç.	0.31	0.13	0.26
22	Çoktan Seç.	0.85	0.25	0.25
23	Kısa cevap	0.50	1.00	0.72
24	Kısa cevap	0.69	0.63	0.61
25	Kısa cevap	0.38	0.25	0.31
26	Kısa cevap	0.46	0.63	0.60
27	Kısa cevap	0.62	0.88	0.79
28	Kısa cevap	0.73	0.63	0.58
29	Kısa cevap	0.46	0.38	0.39
30	Kısa cevap	0.69	0.63	0.48

Okuma Sınavı V2'nin analizlerine göre en başarılı soru tipi kısa cevap soruları olmuştur. Bu sorular hem kolaylık derecesi bakımından (soru 25 hariç), hem ayırıcılık hem de nokta iki serili korelasyon değerleri açısından doyurucudur.

Çoktan seçmeli sorularda, kolaylık derecesi beklenen değerlerin dışında olan maddeler çoğunlukla ayırıcılık ve nokta iki serili korelasyon değerlerinde de istenen değerlerin dışında kalmışlardır.

Sınav güvenirliği ölçeğinde, soru tipleri ayrı ayrı değerlendirildiğinde, eşleştirme sorularında Cronbach alpha değeri 0.67, arayarak okuma (kısa cevap) sorularında bu değer 0.79 çıkmıştır. Ancak çoktan seçmeli sorular farklı tipte okuma gerektirdiğinden (cümle bazlı veya tüm metini okuma) alpha değeri düşük çıkmıştır (0.27).

TARTIŞMA VE SONUÇ

Araştırma sorusu 1

Bu araştırma sorusu, okuma sınavının bağlam geçerliğini göstermek üzere kanıt üretmek amacındaydı. Khalifa ve Weir (2009) adayların okuma sınavı performanslarının sınav dışında genelleştirilebilmesi için gerçek hayatta kullanılan okuma parametrelerinin detaylı olarak tanımlanması ve sınava uygulanması gerektiğini söyler. Aksi durumda Messick'in (1995) geçerlik teorisinde özellikle dikkat çektiği iki problem sıklıkla yaşanabilir. Bunlar *içerik uygunluğu* (content relevance) ve *içeriğin temsil edebilirliğidir* (content representativeness). Bu sınavda içerik uygunluğuna bağıntılı alandaki okuma gereksinimleri incelenerek ve kaynak yazındaki akademik okuma tanımlamaları arasından ODTÜ bağlamına uyan okumaya uygun bir model kullanılarak ulaşılmıştır. İçeriğin temsil edebilir olması ise özellikle fakültelerde görev yapan öğretim elemanlarının önemli olduğunu belirttikleri okuma görevlerinin sınava dahil edilmesiyle sağlanmıştır.

Bu sınavın üretilmesindeki yaklaşım, bir çerçeve kullanılarak sınav içeriği ve parametrelerinin belirlenmesi ve sınav tanımlamalarının bir belge olarak yayınlanmasının alana katkısı büyüktür. Kavramsal düzeyde ele alınan bir becerinin, okuma becerisinin, bir çerçeve kullanılarak ölçülebilir özellikleri belirlenmiş böylece kuram ve veri arasındaki ilişki açıklanmıştır.

İkinci olarak, bu çalışma sınav dizaynı ve geliştirmede olması gereken standartları ortaya koymuştur. Özellikle yüksek etkili sınavlarda sınav geliştiricilerinin sorumluluğunda olan yeterlik kanıtı üretme ve sunma çalışmasının nasıl yapılması gerektiğini göstermiştir.

Son olarak, geçerleme çalışmaları dil öğretimini de etkilemektedir. Sınama ve öğretme arasında doğrudan ve etkileşimli bir ilişki vardır. Kuramsal temeli sağlam, ve bağlamsal geçerliği olan sınavlar, dil öğretiminde öğretmenlere yol gösterirler. Sınav parametreleri dil sınıflarında performans kriterleri olarak kullanılabilir.

Araştırma sorusu 2

Geriye dönük inceleme

Okuma Sınavı V1'de çoğunlukla genel, daha az sıklıkla yerel okuma stratejilerinin kullanıldığı görülmüştür. Hızlı okuma stratejisi gerektirdiği düşünülen soru tipleri (eşleştirme) bu işlevi yerine getirememiştir. Bu yüzden sınavın sonraki versiyonunda hızlı okumayı sağlayacak arayarak okuma soru tipi sınava eklenmiştir.

Okuma sınavının bu versiyonunda katılımcılardan elde edilen okuma tiplerinin kullanımları okuma modeline uyarlanmıştır.

Tablo 6 Okuma modeline uyarlama

		Sıklık
Okuyucu amacı	Dikkatli okuma – Genel ve yerel	9.08
	Hızlı okuma - Göz gezdirme	4.77
	Hızlı okuma - Tarama	0.92
	Hızlı okuma – Arayarak okuma	2.21
Bilişsel işlemler	Metnin zihinde tam modelini oluşturmak	1.94
	Metnin temsilini oluşturmak	3.79
	Cümle bazında anlamak	1.23
Bilgi kaynakları	Metin yapısal bilgisi	0.60
	Genel bilgi / Konu bilgisi	0.79
	Sözdizimsel bilgi	0.63
	Kelime bilgisi	1.77

Buna göre dikkatli okuma en çok kullanılan okuma tipidir. Ardından göz gezdirme gelmektedir. En az kullanılan okuma tipi taramadır.

Bu sınavın ortaya çıkardığı bilişsel işlemler şunlardır: sorulara cevap verirken hem aşağıdan yukarı hem de yukarıdan aşağı okuma süreçleri işlemiştir. TİB'de üst kurlarda (Upper-intermediate ve Advanced) eğitim alan katılımcıların yukarıdan aşağı okuma süreçlerini daha çok kullandıkları anlaşılmıştır. Katılımcılar daha çok metnin zihinde temsilini oluşturmaya çalışmışlardır.

Bilgi kaynaklarında en çok kelime bilgisi kullanılmış, diğer bilgi kaynakları (yapısal, genel, ve sözdizimsel) yaklaşık olarak eşit düzeylerde kullanılmıştır.

İçe dönük inceleme

İçe dönük incelemede soru tipleri etkinleştirdikleri okuma biçimleri ile ayrı ayrı ele alınmıştır.

Kelime soruları dikkatli yerel okuma soruları olarak dizayn edilmiştir. Sonuçlar bu sorularda hem yerel hem de genel dikkatli genel okuma stratejilerinin kullanıldığını göstermiştir. Bu sorular beklenildiği üzere ağırlıklı olarak yerel okuma ile cevaplanmamış, çoğunlukla genel dikkatli okuma stratejileri kullanılmıştır. Aynı zamanda seçenekleri eleme stratejisi de kullanılmıştır. Bu strateji daha çok düşük puan alan katılımcılar tarafından kullanılmıştır. Yüksek not alan kullanıcılar bu sorularda diğerlerine oranla iki kat daha fazla doğru cevap vermişlerdir. Genel olarak kelime soruları genel anlama sorularıyla benzer süreçleri etkinleştirmiştir. Bu yüzden sınavın revizyonunda kelime bilgisini ayrı bir bölümde ve farklı bir şekilde ölçmeye karar verilmiştir.

Eşleştirme soruları hızlı okuma – göz gezdirme yöntemi kullanarak çözülmesi beklenen sorulardı. Bu tür okumada okuyucu metin yapısı bilgisi ve arka plan bilgisini kullanır (Weir ve diğerleri, 2000).

Eşleştirme sorularında düşük not alan öğrencilerin hızlı okuma yerine dikkatli genel okuma yapmaları metin yapısı bilgilerinin az olması veya olmaması, veya arka plan bilgilerini etkinleştirememelerinden dolayı olabilir.

Düşük not alan katılımcılarla yüksek not alan katılımcıların kıyaslanması sonucu bu iki grubun farklı stratejiler uyguladıkları anlaşılmıştır. Yüksek not alan katılımcılar hem hızlı hem de dikkatli okuma stratejileri uygulamışlar, düşük not alan katılımcılarsa sıklıkla sorudan anahtar kelime seçimi yaparak bu kelimeleri metnin için aramayı tercih etmişlerdir.

Genel anlama soruları hem dikkatli genel okuma hem de dikkatli yerel okuma stratejilerini sınamak üzere hazırlanmıştır. Bu sorulardan üç tanesi yerel okumaya odaklanmış, on tanesi bir paragraftan anlaşılacak şekilde hazırlanmış, bir tanesi de metnin bütününden cevaplanacak şekilde hazırlanmıştır.

Bulgular, genel anlama sorularının gerçek hayatta akademik okumayla benzer şekilde bilişsel süreçleri etkinleştirdiğini göstermiştir. Katılımcılar, aşağıdan yukarı ve yukarıdan aşağı olmak üzere her iki tipte okuma sürecini etkileşimli olarak kullanmıştır.

Arayarak okuma soruları katılımcıların uzun bir metnin içinde (2500-3000 kelime) belli bir konunun yerini arayıp bulduktan sonra o konuda derinlemesine okuma yapmalarını sağlamaya yönelik hazırlanmıştır. Bu tür sorularda, özellikle düşük not alan katılımcıların kullandıkları temel stratejilerden biri soruyu okuma ve metni bir süre okuduktan sonra tekrar soruya dönmek idi. Weir et al. (2000) bunun Çin'de uyguladıkları bir sınavda aynı şekilde gerçekleştiğini söylüyor: Okuyucular soruyu iyice anlamadan metinde nasıl arama yapacaklarını bilmediklerinden sık sık soruya dönerek anladıklarını pekiştirmeye çalışıyorlar.

Bu sınavda gözlemlenen modelde, yüksek not alan katılımcılar önce sorudan başlayıp birden fazla stratejiyi kullanarak doğru cevaba ulaşmaya çalıştılar. Düşük not alan katılımcılar daha çok tarama stratejileri kullandılar.

Araştırma sorusu 3

Üçüncü araştırma sorusu madde parametrelerinin sınavın geçerlik savlarını ne derece desteklediğini göstermeyi hedeflemiştir. Buna göre Okuma Sınavı V2'nin sonuçları aşağıda verilmiştir.

Madde kolaylığı parametrelerine göre soruların %63'ü istenilen nitelikte zorluk derecesine sahiptir. Bu sınav ölçüt bağımlı sınav olduğundan, zorluk derecesi 0.40 –

0.80 arasında kabul edilmiş ve buna göre 7 soru 0.80’den büyük, 4 soru ise 0.40’tan küçük çıkmıştır.

Ayırıcılık göstergeleri açısından 19 maddede ayırıcılık beklenen düzeyde iken 11 maddede yeterli ayırıcılık göstergesi bulunamamıştır. Bunun sebeplerinden biri, bu sınavın ölçüt bağımlı sınav olması olabilir (Hambleton ve Novick, 1973).

Genelde, maddelerin yarıdan çoğunda belirlenmiş kolaylık, ayırıcılık ve güvenirlik parametrelerine uygun değerler bulunmuştur. Maddelerin beklenenden daha yüksek IF değeri olması ayırıcılık ve güvenirlik değerlerini de etkilemektedir. Sınırların dışında kalan fazla kolay veya fazla zor görünen maddelerin zorluk dereceleri sınırlar içine çekildiğinde diğer parametrelerde düzelme olacağı düşünülmektedir.

Sonuç

Okuma Sınavı V2, bağlam geçerliği, bilişsel geçerlik ve notlama geçerliği kriterleri temel alındığında başarılı bir sonuç göstermiştir. Sınavın tanımlama dosyası kaynak dizin ve öğrenci gereksinimleri temel alınarak hazırlanmıştır. Bu sınav tanımlama dosyası kullanılarak sınav hazırlanmış, madde parametreleri ve sınavı yapanların bilişsel süreçleri incelenerek ikinci versiyon hazırlanmış ve tekrar madde parametreleri ve bilişsel süreçler incelenerek sınavın ikinci versiyonunun başarılı bir okuma sınavı olduğu görülmüştür.

Çıkarımlar

Kuramsal anlamda sınavın alana ve bilgi birikimine katkısı vardır. Sınavı oluştururken kullanılan sosyal-bilişsel çerçeve ve okuma modeli İngilizce’nin ana dil olarak kullanıldığı bir bağlamda oluşturulmuştur. Bu çalışma bu çerçeve ve modelin İngilizce’nin yabancı dil olarak kullanıldığı bir bağlamda geçerliğini ispatlamıştır.

Okuma modelinde geleneksel okumanın dışına çıkarak hızlı okumanın da (özellikle arayarak okuma) sınanması, sınavın bağlam geçerliğine katkıda bulunmuştur. Bu anlamda Urquhart ve Weir (1998) tarafından varsayılan okuma modelinin anlamlı olduğu ve akademik okumada kullanılabileceği gösterilmiştir.

Pratik anlamda bu sınavın geliştirilme yöntemi diğer sınav geliştiriciler için önemli ipuçları barındırmaktadır. Öncelikle, sınav geliştirmede sistematik bir yöntemin seçilip uygulanması esastır. Bu çalışma Weir'in (2005) sosyal-bilişsel modelinin akademik becerilerin sınanmasında uygulanabileceğini göstermiştir. İkinci olarak, sınav geliştirmede her bir adımda yapılan işlemlerin uygunluğu ve doğruluğu için araştırma yapmak bir sonraki aşamaya doğru bilgilerle geçilebilmesini sağlamıştır. Bu yöntemler hazırlanan sınav, sonuçları itibarıyla kanıtlar gösterilerek savunulabilecektir. Bu sınavın sonuçlarına bakarak verilen kararlar ya da ileri dönük çıkarımların geçerliği rahatlıkla ilan edilebilir.

Bir okuma sınavının bir kuramsal model üzerinden tanımlanması dil eğitimi alanında da olumlu sonuçlar doğurur. Okumanın ODTÜ öğrencisi için işlevselleştirilmesi ve ölçülebilir bir davranış biçimi haline dönüşmesi, okumanın öğretiminde de kullanılabilir. Sınav sonuçları öğrencilerin kuvvetli ve zayıf oldukları okuma alanları hakkında doğru bilgi vereceği gibi, bu bilgiler öğrenciye dönüt vermek için kullanılabilir. Ayrıca, sınavda kullanılan kriterler eğitim ve öğretim için de kullanılabilir. Böylece öğrenciler ihtiyaca yönelik ve eksiksiz bir eğitim olanağına kavuşurlar.

Çalışmanın sınırlılıkları

Örnekleme: Okuma Sınavı V2'nin katılımcıları rastgele örnekleme değil kota örneklemesi yoluyla seçilmiştir. Bu yüzden taraflı bir örneklemedir.

Veri toplama: Hem geri dönük bakış hem de içe bakış yöntemleri özbildirim olduğundan araştırmacı katılımcıların dürüstlüğüne ve becerisine güvenmek durumundadır. Bu çalışmada katılımcıların öznel yargılarının da veriyi etkilemiş olabileceği dikkate alınmalıdır. Ayrıca özbildirim verisi ancak sıralamaya tabii tutulabilecek bir veri tipidir, bu yüzden de çıkarımsal istatistik analizlerinde kullanılamaz.

Veri analizi: Okuma Sınavı V2 verisinin yazıya geçirilmesi ve kodlanması benim tarafımdan yapılmıştır. Kaynak dizin bu tarz çalışmalarda ikinci bir araştırmacının da aynı analizi yapmasını önermektedir. Bu haliyle araştırmanın ikinci sorusunun sonuçları doğrulanabilir nitelikte değildir.

Önerilen araştırmalar

Bu çalışmada bir sosyal-bilişsel çerçeve kullanılmış ve bu bağlam için geçerliği teyit edilmiştir. Aynı çerçevenin farklı bağlamlarda uygulanması kuramsal yapısını güçlendirecektir.

Bu çalışmada kullanılan Okuma Sınavı V2'nin ODTÜ öğrencisinin akademik okuma gereksinimlerine cevap verdiği görülmüştür. Ancak, sınavın kriter-bazlı geçerliği açısından sınavı başka (geçerlik çalışmaları yapılmış) başka dış sınavlarla beraber vererek kriter bazlı geçerliği sağlamak gerekir.

Aynı zamanda bu sınavın Ortak Avrupa Dil Çerçevesine uygunluğunu sınamak için araştırma yapılması önerilir.

ODTÜ-İYS'de okuma dışında yazma ve dinleme sınavları da verilmektedir. Benzer bir çalışmanın diğer iki beceri içinde yapılması önerilir.

APPENDIX L: TEZ İZİN FORMU FORMU / THESIS PERMISSION FORM

ENSTİTÜ / INSTITUTE

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences □

Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences ■

Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics □

Enformatik Enstitüsü / Graduate School of Informatics □

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences □

YAZARIN / AUTHOR

Soyadı / Surname : Akşit
Adı / Name : Zeynep
Bölümü / Department : İngiliz Dili Öğretimi

TEZİN ADI / TITLE OF THE THESIS (**İngilizce** / English) : Validating aspects of a reading test

TEZİN TÜRÜ / DEGREE: **Yüksek Lisans** / Master □ **Doktora** / PhD ■

1. **Tezin tamamı dünya çapında erişime açılacaktır. /** Release the entire work immediately for access worldwide. □

2. **Tez iki yıl süreyle erişime kapalı olacaktır.** / Secure the entire work for patent and/or proprietary purposes for a period of **two year. *** □

3. **Tez altı ay süreyle erişime kapalı olacaktır.** / Secure the entire work for period of **six months**. * ■

* *Enstitü Yönetim Kurulu Kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir.*
A copy of the Decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.

Yazarın imzası / Signature .... **Tarih** / Date