THE EFFECT OF ITEM FORMAT ON THE CHOICE OF READING AND TEST-TAKING STRATEGIES

HATİCE AKGÜN

BOĞAZİÇİ UNIVERSITY

2018

THE EFFECT OF ITEM FORMAT ON THE CHOICE OF READING AND TEST-TAKING STRATEGIES

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfilment of the requirements for the degree of

Master of Arts

in

English Language Education

by

Hatice Akgün

Boğaziçi University

2018

DECLARATION OF ORIGINALITY

I, Hatice Akgün, certify that

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

Signature.....

Date ....26..03..2018.....

ABSTRACT

The Effect of Item Format on the Choice of Reading and Test-Taking Strategies

Investigating the processes that test takers undergo when they answer reading comprehension questions is of utmost importance to make claims about cognitive validity of a test. As such, this study was motivated by the need to explore reading processes, reading and test-taking strategies that test takers utilize when they take a test in different formats, namely multiple-choice and open-ended formats. In order to find out whether there were any differences in terms of reading processes and strategy use in these question formats, the data were collected through a triangulation of eye tracking technology, retrospective verbal reports and short semi-structured interviews. The results showed that the scores of open-ended questions were higher than their multiple-choice equivalents. Eye-tracking data showed that the percentage of careful reading was higher in open-ended test format and a further analysis on eye movement data showed that test takers spent a longer time in the interest areas while answering open-ended questions. The results of verbal report data displayed that test takers used test-taking strategies more in multiple-choice items while they use reading strategies more in open-ended items. Lastly, interview results showed that the majority of test takers reported to have comprehended the text in open-ended format more. This study is important in terms of proving that test format has an effect on the reading processes of test takers and multiple-choice format reduces the amount of careful reading and alters normal reading processes. The study also shows the promising and valuable contribution of eye-tracking technology in investigating cognitive validity.

ÖZET

Soru Formatının Okuma ve Test Çözme Stratejilerinin Seçimi Üzerindeki Etkisi

Okuma becerilerini ölçen soruları cevaplarken, öğrencilerin geçirdikleri süreçleri araştırmak sınavın bilişsel geçerliliğini ispatlamak açısından büyük önem taşımaktadır. Bu amaçla, bu çalışma, öğrencilerin farklı formatlarda – açık uçlu ve çoktan seçmeli sorularda- bir sınav çözerken faydalandıkları okuma süreçleri, okuma ve test çözme stratejilerini incelemeyi hedeflemektedir. Bu soru formatlarındaki okuma süreçleri ve strateji kullanımındaki farklılıkları öğrenmek amacıyla, göz hareketlerini izleme, geriye dönük sözlü raporlama ve yarı yapılandırılmış görüşmelerden oluşan veri üçlemesiyle çalışma için gereken veri toplanmıştır. Sonuçlar, ilk olarak açık uçlu sorularda öğrencilerin başarısının arttığını göstermiştir. Göz hareketlerini izleme sonucu olarak, öğrencilerin açık uçlu sorularda dikkatli okuma oranlarının daha yüksek olduğu görülürken, daha ileri bir analiz öğrencilerin, açık uçlu sorularda, okuma parçasında soruyla ilgili olan alanlarda daha fazla zaman geçirdiklerini gösterdi. Sözlü raporlama sonuçları ise, öğrencilerin çoktan seçmeli sorularda test çözme stratejilerini daha çok kullandığını ancak açık uçlu sorularda ise okuma stratejilerini daha çok kullandıklarını gösterdi. Görüşmelerde ise, testi çözen öğrencilerin çoğunluğu açık uçlu sorularda ilgili okuma parçasını daha iyi okuduklarını ve anladıklarını ifade ettiler. Bu çalışma öncelikle, test formatının okuma süreçleri üzerinde etkili olduğunu, çoktan seçmeli sorularda dikkatli okuma süresinin azaldığını ve normal okuma süreçlerinin değiştiğini göstermesi açısından önemlidir. Ayrıca, göz hareketlerini izlemenin, bir testin bilişsel geçerliliğini kanıtlamadaki önemli katkılarını göstermektedir.

ACKNOWLEDGEMENTS

This thesis could not have been complete without the involvement and contribution of many people and I am glad to have this opportunity to express my gratitude to those who helped me complete it. First of all, I would like to express my profound gratitude and admiration to my advisor, Assist. Prof. Aylin Ünaldı for her guidance, care and continuous encouragement. Without her expertise and meticulous comments, I would not have been able to finish my thesis. I would also like to thank the other members of my committee, Assist. Prof. Senem Yıldız and Assist. Prof. Mustafa Polat for their illuminating comments and suggestions on this study along with their warm encouragement. I would particularly like to thank Assist. Prof. İnci Ayhan for giving me the opportunity to conduct the eye-tracking experiments at Boğaziçi University Vision Lab and I am also indebted to Emre Oral who contributed greatly to the implementation of eye-tracking in the study.

I would also like to thank my family members and friends for their support. I dedicate all my successes to my parents, Nilgül and Siraç Evcil, as they have always kept encouraging me all the way. And most of all, my special thanks go to my best friend and my husband, İlkay Akgün, for always being with me through the entire journey of my MA study. I sincerely thank my sister, Melike Evcil, and my brother, Hakan Evcil, for the continuous support and motivation they provided. I also thank my dearest friends, Yeliz Ergül Özcan, Beyza Öz Dal and Burcu Kayarkaya for their precious friendship and support. Last but not least, I should thank my little one, Çınar Ali Akgün, who made the process of thesis writing more challenging for me, but at the same time more satisfying. I am deeply grateful to him for being the source of joy and courage in my life at all the times I lost faith in myself while writing this thesis.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1
1.1 Theoretical background 1
1.2 Motivation for the study 4
1.3 Significance of the study 5
CHAPTER 2: REVIEW OF THE LITERATURE 8
2.1 Test validation 8
2.2 Models of second language reading 12
2.3 Strategies in L2 reading tests 20
2.4 Test-format effect 26
2.5 The use of verbal reports 30
2.6 Eye-tracking technology 36
2.7 Conclusion 47
CHAPTER 3: METHODOLOGY 49
3.1 Introduction 49
3.2 Research questions 49
3.3 Participants 50
3.4 Instruments 52
3.5 Procedure 57
3.6 Data analysis 61
3.7 Conclusion 65
CHAPTER 4: RESULTS 66
4.1 Introduction 66
4.2 The effect of test method on test performance 66
4.3 The comparison of eye movement data in both formats 68
4.4 The comparison of reading and test-taking strategy use depending on test format 73
4.5 Test takers' own perceptions about test method effect 82
4.6 Conclusion 85
CHAPTER 5: DISCUSSION 86
5.1 Introduction 86
5.2 The effect of test method on reading performance 87
5.3 The results of eye-tracking data 88
5.4 The results of verbal report data and semi-structured interviews 91

5.5 Item by item comparison between eye movement and verbal report data to investigate cognitive processes depending on format effect 95
5.6 Conclusion 100
CHAPTER 6: CONCLUSIONS 101
APPENDIX A: CONTEXTUAL PARAMETERS: ITEM-LEVEL 105
APPENDIX B: CONTEXTUAL PARAMETERS: TEXT-LEVEL 106
APPENDIX C: TEXT 1 MULTIPLE-CHOICE FORMAT……107
APPENDIX D: TEXT 2 MULTIPLE-CHOICE FORMAT 108
APPENDIX E: TEXT 1 OPEN-ENDED FORMAT 109
APPENDIX F: TEXT 2 OPEN-ENDED FORMAT 110
APPENDIX G: INFORMED CONSENT FORM 111
APPENDIX H: AREAS OF INTEREST 113
APPENDIX I: READING STRATEGIES CODING RUBRIC 114
APPENDIX J: TEST-MANAGEMENT STRATEGIES CODING RUBRIC 115
APPENDIX K: TEST-WISENESS STRATEGIES CODING RUBRIC 117
APPENDIX L: EYE-TRACKING STATISTICS FOR CORRECT RESPONSES.118
APPENDIX M: RESULTS OF T-TEST FOR READING STRATEGIES…….121
APPENDIX N: RESULTS OF T-TEST FOR TEST-TAKING STRATEGIES…..122
REFERENCES 123

LIST OF TABLES

Table 1. Facets of Validity as a Progressive Matrix……9

Table 2. Text Analysis of Reading Materials Used in the Study……...53

Table 3. Reliability Statistics for Text 1: Origins of Cetaceans……....55

Table 4. Reliability Statistics for Text 2: Swimming Machine……..…...55

Table 5. Items of Text 1 MC & OE and Text 2 MC & OE……...56

Table 6. Procedure of the Study……61

Table 7. Descriptive Statistics for Method 1 and Method 2……..67

Table 8. Descriptive Statistics for Text 1 and Text 2……67

Table 9. The Results of Two-way ANOVA……..68

Table 10. Eye-Tracking Statistics for Text-based Careful Reading, TRT and TFC..69

Table 11. Eye -Tracking Statistics for Item 1……....70

Table 12. Eye-Tracking Statistics for Item 2……70

Table 13. Eye-Tracking Statistics for Item 3……71

Table 14. Eye-Tracking Statistics for Item 4……71

Table 15. Eye-Tracking Statistics for Item 5……72

Table 16. Eye-Tracking Statistics for Item 6……72

Table 17. The Summary Table for Overall Item-based Differences……73

Table 18. The Summary Table for Overall Differences for Correct Respondents…74

Table 19. The Frequency of Overall Reading Strategies……...75

Table 20. The Frequencies of Test-Management Strategies……77

Table 21. The Frequencies of Test-Wiseness Strategies……...77

Table 22. Option- Related Strategies……78

Table 23. Item 1- Strategy Use ……....79

Table 24. Item 2- Strategy Use ……80

Table 25. Item 3- Strategy Use…….80

Table 26. Item 4- Strategy Use…….81

Table 27. Item 5- Strategy Use ……81

Table 28. Item 6- Strategy Use ……82

Table 29. Results of Interviews…….…...83

LIST OF APPENDIX TABLES

Table H1. Areas of Interest for Text 1……113

Table H2. Areas of Interest for Text 2……113

Table L1. Eye- Tracking Statistics for Item 1……..118

Table L2. Eye- Tracking Statistics for Item 2……..118

Table L3. Eye- Tracking Statistics for Item 3……..119

Table L4. Eye- Tracking Statistics for Item 4……..119

Table L5. Eye- Tracking Statistics for Item 5……..120

Table L6. Eye- Tracking Statistics for Item 6……..120

Table M1. Descriptive Statistics for Reading Strategies……..121

Table M2. T-Test Results for Reading Strategies……121

Table N1. Descriptive Statistics for Test-Taking Strategies……122

Table N2. T-Test Results for Test-Taking Strategies……...122

CHAPTER 1

INTRODUCTION

1.1 Theoretical background

It is widely accepted that test development plays a worldwide importance considering the fact that test results are seen as gatekeepers in many societies to be accepted to most programs (Shohamy, 2001). The fact that tests results have such a crucial effect on the decisions made by the institutions and on the lives of all stakeholders makes test validation a serious concern for test developers. What is meant by test validation is investigating whether a test measures what it intends to measure. While it has such a simple definition, there have been many diverse attempts to formulate an efficient test validation process. Dissatisfaction with former test validation methods which were based on more outcome-based approaches resulted in more process-based approaches to validation. As it is emphasized in Wu and Stone (2016), the previous product-based approaches were based on the scores from a test and investigated simple correlations with other outcome measures as evidence of construct validity and/or criterion validity (predictive or concurrent validity). However, such outcome-based methods were limited and the numbers couldn't convey conceptual information and test scores couldn't explain how test takers derived their answers (Weir, 2005). Therefore, the crucial point is not what an item is thought to be testing, but what kind of processes are triggered by the correct responses and whether these cognitive processes are the intended ones (Alderson,2000). Furthermore, it is also pointed out by Field (2011, 2012) that it is necessary to find out whether the mental processes elicited from a test taker by an item resembles the processes that s/he would employ in non-test conditions. In this sense, it is quite obvious that finding out cognitive processes triggered by an item

and the processes that a test taker undergoes while responding to an item are of utmost importance in order to make validity claims. Hence, test takers' cognitive processing and their perceptions along with the skills and strategies they use have become substantial analyses.

As it is a well-known fact, reading comprehension is a complex process involving a variety of skills and strategies such as skimming, scanning, search reading or careful reading at local or global levels (Urquhart and Weir, 1998; Khalifa and Weir, 2009). In addition to these, there are also a variety of metacognitive strategies that test takers can employ. The first group is planning strategies in which readers set a goal on how to read. The second is monitoring strategies in which comprehension is checked at different stages of the reading task by the reader and the last one is evaluation strategies in which readers remediate whenever necessary or evaluate/assess their performance in reading comprehension (Purpura, 1999; Khalifa and Weir, 2009; Pressley and Afflerbach, 1995). Furthermore, Cohen (2006) suggests that test-taking strategies can be used for validation purposes, too. As to the test-management strategies, Cohen (2006; 2007) divides them into two categories as test-management strategies and test-wiseness strategies. While test-management strategies are used by the test taker to respond meaningfully to test items and tasks, test-wiseness strategies are defined as strategies in which test takers use the knowledge of testing formats and other peripheral information to answer test items without going through the intended processes (Cohen, 2006). In this context, investigating the test-taking strategies used by the test taker is important in that it can shed a light into understanding how a test taker responds to an item and whether the strategies that are used are relevant to the intended processes in the construct. Therefore, as it is pointed out by Wu and Stone (2016), test-wiseness strategies might

have contributed to a test taker's score and this contribution is totally irrelevant to the construct. Finding out these kinds of construct-irrelevant strategies will be crucial in revising test items and very helpful for test developers.

At this point, it is important to note that test takers' use of test-taking strategies is induced by test format effect (Sarnaki, 1979) and therefore understanding the processes triggered by each item format, for example open-ended and multiple-choice question types, in reading tasks plays a crucial role in validating the test. As Lim (2014) highlights, to come up with the correct answers, test-takers employ some extra mental processes, or item-responding processes along with the genuine reading processes. Hence, different test formats can tap into different aspects of the construct and they can also involve construct-irrelevant elements. For example, it was shown in Rupp, Ferne &Choi (2006) that test takers approached multiple-choice tasks as a problem-solving task rather than as reading comprehension and they scanned the texts for key word matching instead of reading carefully. In addition, it is also clear that open-ended and multiple-choice formats differ in terms of their cognitive demands (Martinez, 1999; Rauch and Hartig, 2010; Ozuru, Briner, Kurby, McNamara, 2013) and test-taking strategies specific to the multiple-choice format can undermine a test's cognitive validity (Field, 2011).

In order to explore these cognitive processes and reading and test-taking strategies, research has been mostly done via retrospective and concurrent verbal reporting by test-takers about what they think when they answer each question (Bax and Weir, 2012). While verbal reports are mainly viewed as valuable and reliable source of information about cognitive processes, there are also some concerns about their use to investigate cognitive processes. Some of these concerns are the reactive effects of verbal report such that on-line reporting may change the original thought

processes and the data may be contaminated by the shared assumptions or perceptions (Cohen, 1998). As to the retrospective think-aloud, it is suggested that during verbal reporting, test-takers may report false processes, add unnecessary information to rationalize the process they have undergone or omit some parts that should have been included (Green, 1998). Due to these concerns, Cohen (1998) suggested that verbal report should be used as a complementation rather than a replacement for other means of research. Eye-tracking research appears as a complementary method to retrospective verbal reports as it provides additional opportunities to investigate readers' actual behaviour and a greater insight into the probable cognitive processes (Bax and Weir, 2012). Rayner (1998) claims that eye movement data reflect moment-to-moment cognitive processes and this method doesn't interfere with the reading process in any way (Rayner and Sereno, 1994).

1.2 Motivation for the study

Although research on the cognitive processes that certain item types trigger during reading tests is not scarce, a direct comparison of multiple-choice and open-ended items with triangulated methods such as think-aloud and eye-tracking is still to be done. The study which makes use of similar methods and which also investigates test format effect through stem-equivalent versions of two reading text formats is by Lim (2014). However, it should be noted that the primary aim of that study is to validate TOEFL iBT and the study implemented stimulated recall interview and strategy questionnaire only to those participants who took the tests in multiple-choice format, not the ones who took the tests in open-ended format. In addition to this, stimulated recall interviews were administered only on three items in the test. Therefore, this implies the need to compare open-ended and multiple-choice item types in a more

systematic way by collecting data systematically on each item through the triangulation of eye movement data and verbal reports. Therefore, the current study was first motivated by this gap in research and also from my personal experiences and impressions as a test developer. To this end, in this study, to understand the cognitive processes that. test takers engage in while answering reading comprehension questions in two different test formats (multiple-choice and open-ended formats), reading and test-taking strategies that they utilize will be investigated with the help of eye-tracking data and immediate retrospective verbal report.

1.3 Significance of the study

Investigating cognitive processes that test takers go through while responding to a test item is of utmost importance as it can provide test developers with a deep insight as to the interpretation of a test's validity. In this way, the process by which test takers come up with the correct answers can be understood and this process can be compared with the processes that test writers intended while developing the test items. In the same vein, test format, namely open-ended and multiple-choice, implemented to measure reading comprehension may affect the usual reading processes as the format may require some extra processes that are irrelevant to intended processes of an item or even the same item prepared in different formats may alter how much of the text is processed by the test taker to find the correct answer. In addition, test taker can utilize test-taking strategies to cope with the item demands and find the correct answers. Exploring these test-taking strategies can show whether these strategies are irrelevant to the reading construct. As in some cases, it might be possible that test takers can choose the correct answers, not by

going through the intended reading processes but rather by making use of some clues or other problem-solving abilities and test format can be seen as the cause of this problem to some extent.

In several studies, to explore what test takers do when they respond to questions, their online eye movements are recorded to find out how much of the text is processed by respondents in each item depending on the format of the item, what reading types are used by them and whether these reading types are the intended ones by test developers, how much time they spent on processing relevant areas. Eye-tracking tool is advantageous as it doesn't interfere with the respondents' reading processes and it isn't expected to disrupt the usual process. To this end, it is seen to be helpful in revealing the differences evoked by two different test formats. In addition to recording eye movements, respondents in this study were asked to give an immediate verbal report about how they answered each question and report the processes that they went through. These two methods complement each other in order to investigate both observable and non-observable processes and produce validity evidence for each test format.

To this end, this study may yield significant results owing to three different reasons. First, exploring test format effects in a controlled way by creating equivalent items in terms of content and stem can facilitate making good comparisons between these two formats. Secondly, the triangulation in the methodology by making use of eye-tracking technology, retrospective verbal reports and short semi-structured interviews with participants can bring about a deeper insight into the possible effects of item format differences in reading behaviour and comprehension and can result in valuable information as to which item format can be a more valid measure of the reading construct and which one includes more irrelevant variables to the construct.

Thirdly, the use of eye-tracking in exploring test format effect in reading tests is limited in the research studies and it is hoped that the findings of this study will both contribute to the discussion of test format effect and also inspire further research in this field.

The following chapters are organized as follows: Chapter 2 will explain test validation, models of second language reading and strategies used in L2 reading tests. Chapter 3 will focus on the methodology and the design of the current study. Chapter 4 will present the results of the study and in Chapter 5, these results will be discussed. In Chapter 6 the conclusions drawn from the study will be reported along with the implications and the limitations of the study.

CHAPTER 2

REVIEW OF THE LITERATURE

2.1 Test validation

Validity in testing and assessment has traditionally been defined as finding out whether a test "measures accurately what it is intended to measure" (Hughes, 1989; 22). Despite being a very traditional definition, it is both comprehensive and essential to understand the basic ideas underlying validity (Lim,2014). In its early days of investigation, validity was divided into three separate parts and these were categorized as criterion-oriented validity (predictive validity and concurrent validity), content validity and construct validity (Cronbach and Meehl, 1955). Criterion-oriented validity can be defined as forming a relationship between a particular test and a criterion to which predictions are to be made. Content validity is the attempt to show if the content of a test is a representative sample of the domain that is to be tested. As the construct is a hypothetical rather than observable entity, construct must be defined in an operational way and measured by linking it to something observable (Fulcher and Davidson, 2007). Therefore, construct validation is described as the extent to which performance on tests is consistent with predictions that are made on the basis of a theory of abilities, or constructs (Bachman, 1990). What continued to be influential in test validation from the categorization of Cronbach and Meehl (1955) was the importance of construct validity (Fulcher and Davidson, 2007). To this end, Messick (1989, 1995) suggested that content and criterion- related evidence are aspects of construct validity. He proposed a unified concept of validity which integrates considerations of content, criteria and consequences into a construct framework for the empirical testing of hypothesis about score meaning and which addresses score meaning and social values in test interpretation and test use. In

Messick's (1980; 1995) unified framework of validity, there is a four-way classification described by two facets which are the source of justification of the testing (evidence or consequence) and the function or outcome of testing (test interpretation or test use). It is emphasized that a salient feature of this framework is that construct validity is an essential part of every cell and to interpret a test score, evidence for construct validity must be gathered and value implications for this interpretation must be considered (Messick, 1995). This framework is illustrated in Table 1. Bachman (1990) also suggested that in test validation, what was validated was not the validity of test content or scores, but the validity of how the information gathered during testing procedure is interpreted or used. To this end, validity of a test or a test score can be claimed by referring to specific abilities the test intends to measure or the uses for which the test is intended (ibid.).

Table 1. Facets of Validity as a Progressive Matrix

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

Messick (1995) also describes two major threats to construct validity which are construct underrepresentation in which the assessment is too narrow and fails to include important dimension (p. 742) and construct-irrelevant variance in which the assessment is too broad containing excess variance associated with other constructs or method variance such as response sets or tendency to guessing which affect the responses in an irrelevant way to the construct. Due to this paramount importance given to construct validity, it is worth mentioning the ways of collecting evidence to support construct validity. Messick (1988; as cited in Bachman, 1990) summarizes

that these ways may include any or all of the following: (1) the examination of patterns of correlations among item scores and test scores with the help of factor analysis, the multitrait-multimethod design and/or experimental evidence; (2) analyses and modelling of the processes underlying test performance; (3) studies of group differences; (4) studies of changes over time or (5) investigation of the effects of experimental treatment. Bachman (1990) underlines the importance of analysing the process of test-taking among all these methods by asserting that studying test-taking processes seems to hold "the most promise for providing new insights into the factors that affect the test performance" (p. 258).

While cognitive validity seems to be a relatively new perspective when compared to the importance given to construct validity (Lim, 2014), the implications about the importance of studying and understanding the cognitive processes of test takers can be traced back to Messick (1988 as cited in Bachman, 1990, p. 269) when he stated the following sentences:

> in numerous applications of … techniques for studying the process, it became clear that different individuals performed the same task in different ways and that even the same individual might perform in a different manner across items or on different occasions… That is, individuals differ consistently in their *strategies and styles of task performance*. (emphasis in original) (Messick 1988, p.54)

This quotation obviously emphasizes the importance of studying cognitive processes that test-takers engage in because these processes can shed light on the claims or interpretations that will be made based on the test scores. Alderson (2000) also underlines the importance of studying the cognitive processes triggered by the test items by asserting:

> [T]he validity of test relates to the interpretation of the correct responses to items, so what matters is not what test constructors believe an item to be testing, but which responses are considered correct and what process underlies them. (p.97)

As it is evident in both quotations, cognitive processes can change depending on the individual test-taker and the process might be quite different in each test item as the cognitive demands triggered by each task are not the same. Therefore, it seems crucial to investigate the cognitive processes in order to make sound validity claims. To this end, Weir (2005) proposes a comprehensive model of validity in which test developers are required to generate evidence of the validity of a test from different aspects. This unified validity framework is defined as "socio-cognitive" framework in that the abilities that are tested are indicated by the mental processing of the test-taker (the cognitive dimension) and the use of language to perform tasks is viewed as a social rather than a purely linguistic phenomenon. The model is divided into "priori (before- the-test event) validation of context and cognitive validity and posteriori (after-the-test event) validation of scoring validity, consequential validity and criterion-related validity" (Khalifa and Weir, 2009, p.4). In this model, test takers and their characteristics (physical/physiological, psychological and experiential) play a fundamental role as they are considered to be relevant items to test design, too. Weir (2005) further asserts that statistical data cannot be self-exploratory in terms of the conceptual labels of constructs and test developers must always be aware of the fact that they need to find out and define what is measured and whether a test is adequate in operation or not.

In addition, Khalifa and Weir (2009, as cited in Weir, Hawkey, Green, Ünaldı, Devi, 2009) point out the limitations of purely quantitative approaches with regard to validation purposes such as factorial approach in that they disregard the

types of reading and the levels of cognitive demand imposed by test tasks. Furthermore, a product-based approach to test validation is criticized on the grounds that it is not based on a sound analysis of salient cognitive processes. The product-based approach is not enough to show what is happening in reality when a reader processes a text in terms of identifying the skills and strategies that contribute to the reading process and they conclude that these can be understood only if cognitive processes are analysed in enough detail. Field (2012) also highlights the importance of cognitive validity by indicating that the mental processes elicited by a task need to resemble the processes that a test taker would employ in contexts beyond the test. Bax (2013) interprets this claim by saying that it is a reinterpretation of Messick's construct-irrelevant variance or construct under-representation from a cognitive processing perspective. Based on all the points raised above, it is clear that validity is an indispensable part of test development and studying the cognitive processes that test-takers go through during a test is of utmost importance before making sound claims about validity (Bax and Weir, 2012; Weir, 2005; Khalifa and Weir, 2009). Such an investigation can only be done, however, when there is a sound theoretical basis supporting the premises of research. To this end, several models of second language reading will be focused on in the next part to understand the expected cognitive processes.

2.2 Models of second language reading

Reading can be defined as a complex set of skills and processes that interact in a complicated manner to produce comprehension (Grabe, 2009) and due to this complexity, there have been numerous attempts to define the nature of reading. As Alderson (2000) remarks, when different aspects of reading contributing to its nature

and the complexities of texts also affecting the process are taken into consideration, it is almost impossible to make a comprehensive overview of all the attempts to explain the nature of reading. Therefore, only two models of second language reading that are relevant to the present study will be reviewed after the presentation of some topic-related background knowledge.

First of all, one should be cautious about the fact that although L1 and L2 reading share some similarities, L2 reading is a more complex process in that it includes different variables such as L1 and L2 orthographic and processing differences, educational and developmental differences (difference in L1 and L2 reading experiences), cultural and institutional variables (Grabe, 2009). There were some attempts to explain the relationship between L1 and L2 reading such as linguistic interdependence hypothesis (Cummins, 1979) which stated that first and second language reading is interdependent and development of one will facilitate the reading in the other. However, this view was challenged by some researchers who suggested that L2 knowledge is a stronger predictor of L2 reading than general L1 literacy skills (Bernhardt and Kamil, 1995). In addition to these views, Grabe and Stoller (2002) present metaphorical models of reading which resulted from comprehension research for over thirty years. These models are bottom-up models, top-down models and interactive models. The authors assert that these models can be thought as an initiation into thinking about reading comprehension, but they cannot clarify more recent research advances. Bottom-up models are described as the ones in which the reader creates a piece-by-piece mental translation of the information in the text, with very little interference from the reader's own background knowledge (Grabe and Stoller, 2002) and these are serial models in which processing happens in a hierarchical way from grapho-phonic, phonemic, syllabic, morphemic, word to

sentence levels right through the text level (Weir, Hawkey, Green, Ünaldı, Devi, 2012). In top-down models, reading is mainly directed by reader goals and expectations. Grabe and Stoller (2002) assert that this model is again metaphorical in that it characterizes "the reader as someone who has a set of expectations about the text information and samples enough from the text to confirm or reject these expectations." (p.32). Authors further assert that inferencing is a feature of top-down models (ibid). In top-down models, readers make use of general and domain specific information to predict text meaning and sentences and words in text (Bernhardt, 1991; as cited in Weir et al., 2012). As a compromise between these two models, interactive models are suggested and, in these models, the main rationale is that one can take useful ideas from a bottom-up view and combine them with the key ideas from a top-down view and to illustrate this view, readers need quick word recognition skills but at the same time, they will need background knowledge to understand the text or inferencing (Grabe and Stoller, 2002). According to the modified interactive models or a hybrid bottom-up/top-down model, readers combine reasonable processes from both top-down and bottom-up models (Weir et al., 2012). Interactive Compensatory Model (Stanovich, 1984, 2000) is an example of such interactive models and it mainly argues that the problems in reading result in increased interaction and compensation even among processes that are expected to operate automatically under normal conditions and this compensatory mechanism helps unskilled readers to compensate by resorting to top-down processes (Khalifa and Weir, 2009). Grabe and Stoller (2002) exemplify this by mentioning the situations in which readers use context clues to understand a text more efficiently or deciding what a word means when their expected abilities break down. Weir et al. (2012) assert that this is a model which underlines the importance of reader role in

reading comprehension. Another model that puts forward that reading is a reader-driven process is Construction-Integration model by Kinstch (1988). The Construction-Integration model developed by Kinstch (1988, 2004) suggests that there are many levels of representation constructed during comprehension three of which are the surface code (decoding the words), the propositional text-base (make meaning from words), the situation model (a mental image is connected to prior knowledge and what is coming is predicted). In this model, a connectionist network is created, modified and updated during comprehension in that as the text is read, sentence by sentence, a set of concept and proposition nodes are activated. Some of these nodes are related to explicit information in the text while others are activated by the world knowledge, rules, etc. from the long-term memory. Therefore, it is possible to claim that reading process in this model combines bottom-up visual information with top-down world knowledge that reader brings to the task (Khalifa and Weir, 2009). Regarding the strategy use, it can be said that strategy is simply a piece of knowledge stored in long-term memory that is periodically activated and recruited during integration. Therefore, strategies do exist in the model but "they don't drive the comprehension engine" (as cited in McNamara, 2007, p. 11). In the subsections below, 'cognitive processing model' by Khalifa and Weir (2009) and 'compensatory theory of second language reading' will be discussed.

2.2.1 Cognitive processing model

The cognitive processing model on reading suggested by Khalifa and Weir (2009) is a synthesis of already existing views from previous research on cognitive processing, componentiality, various types of reading and the various models that have tried to explain reading comprehension. In this comprehensive model, metacognitive activity

of a *goal setter* can be seen on the left column and it is important in deciding what type of reading to employ or the purpose of the reading. With the help of goal setter, crucial decisions are taken which affect the level of processing to be activated in the central core of the model. The various elements of this processing core in the middle column which are affected by the decisions taken in the goal setter can be seen. Next, monitoring, which can be used in each level of processing and remediation where necessary, refers to the ways that readers resort to whenever there is a comprehension breakdown. On the left column, it is possible to find the types of reading that can be selected while responding to a task. The components of knowledge base required to comprehend a text are located in the right column.

When the types and levels of reading suggested in the model are considered, it is obvious that this model is an extension of Urquhart and Weir (1998). As to the levels of reading, comprehension at the local level is described as understanding the propositions at the level of micro-structure such as a sentence or a clause. In the model, authors suggest that local comprehension takes place "at the levels of decoding (word recognition, lexical access and syntactic parsing) and establishing propositional meaning at the sentence and clause level" (Khalifa and Weir, 2009, p.45). With regard to the comprehension at global level, it refers to the understanding propositions beyond the level of micro-structure. Therefore, it includes "any macro-propositions such as main ideas, the links between those macro-propositions and the way in which the micro-propositions elaborate on them" (ibid, p. 45). Careful global reading is described as identifying main ideas by establishing macro-structure, which means understanding how the ideas in the whole text interrelate with each other and the reader's purpose. That is, the reader starts to read from the very beginning of the text and continues until the end and integrates the new information into a mental

model. Search reading at global level happens when the reader tries to gather macro-propositional information quickly and selectively through short cuts. In global comprehension, skimming is the attempt to establish the superordinate macro-proposition (Kong, 1996 as cited in Khalifa and Weir, 2009). Types of reading are divided into two categories which are careful and expeditious reading. In careful reading, the goal is to get complete meanings from the presented material and the approach in this reading type is "slow, careful, linear, incremental reading for comprehension" (Khalifa and Weir, 2009, p.46). Careful local reading involves processing at the decoding level until the basic meaning of a proposition is established and local inferencing may be necessary to build a mental model at the sentence level. Bax and Weir (2012) suggest that processing in careful reading can be at sentential, intersentential, text and multi-text levels. Expeditious reading includes quick, selective and efficient reading in order to find out the desired information. This type of reading consists of skimming, search reading and scanning. Skimming is defined as reading to obtain the gist and general impression. Scanning involves reading selectively in order to achieve very specific reading goals such as looking for specific dates, years, names, words, figures, percentages, etc. As to the search reading, it is defined as reading quickly to locate information on a predetermined topic. It differs from scanning and skimming in that in search reading, the reader doesn't look for exact word matches, but the words in the same semantic field and the reader doesn't have to build a macro-propositional structure for the whole text, but s/he only searches for information on pre-determined topics (Khalifa and Weir, 2009). Understanding expeditious reading is necessary as careful reading models cannot explain how successful readers handle expeditious reading. Rayner and Pollatsek (1989) assert that speed and efficacy of reading are also important

points to consider during comprehension. Weir et al. (2012) suggest as a result of a study on undergraduate students, expeditious reading skills are as critical as careful reading skills with regard to their academic study and expeditious reading is more problematic for both L1 and L2 readers in some cases.

The purpose of reading is crucial in determining the type of reading and components of goal setter and monitoring are metacognitive mechanisms that mediate among various processing skills and knowledge sources. The core processes can be listed as word recognition, lexical access, syntactic parsing, establishing propositional meaning, inferencing, building a mental model, creating a text level representation and creating an intertextual representation. The knowledge sources that a reader can make use of are summarized as text structure knowledge, general and topic knowledge, syntactic knowledge, lexicon in terms of meaning and form (Khalifa and Weir, 2009). The reader may set a goal in terms of the reading type according to the demands of the task and the processing required will be influenced by task demands. In this context, monitoring will help to understand whether the reader can successfully get the writer's arguments and remediate when there is a breakdown in comprehension. This comprehensive reading model, which is both versatile and most up-to-date model will be used as a reference in the current study.

2.2.2 Compensatory theory of L2 reading

The interactive compensatory model suggested by Stanovich (1984) claims that readers resort to some compensatory strategies during reading comprehension breakdowns and to exemplify this, it was claimed that readers with limited word recognition skills are more likely to depend on the context to infer meaning than readers with automatized word recognition skills. In addition, the model suggests that

the use of compensatory strategies cannot distinguish good and bad readers (Lim, 2014).

In addition to this model, Bernhardt (1991, 2005) introduces interactive concepts to create formal models of second language reading. Compensatory model of second language reading reflects Stanovich's (1984) concept of compensatory processing and this L2 compensatory model consists of three components, which are L1 literacy skills, L2 language knowledge and unexplained variance. 20 % from L1 literacy skills and 30% from L2 language knowledge explain second language reading comprehension in this model (Bernhardt, 2005). McNeil (2012) extends the compensatory models of second language reading by adding strategic knowledge and background knowledge to L2 language knowledge and L1 reading ability. Depending on the review of empirical investigations and theoretical explanations, it is claimed that strategic and background knowledge also function as a compensatory resource.

This compensatory model of L2 reading reminds Bachman's (1990) concept of learner's strategic competence that is included the communicative language ability model. In this model, Bachman suggests that communicative language ability is composed of language competence, strategic competence and psychophysiological mechanisms and these interact with the language use context and the user's knowledge structures. In this model, Bachman asserts that language competence interacts with strategic competence and strategic competence consists of the ability to assess, plan and execute appropriate interactional language use by the most effective means. In strategic competence, learners decide on the necessary linguistic resources to achieve a communicative goal, retrieve linguistic elements from language competence and manage neurological and physiological processes to implement a plan. Therefore, Bachman interprets strategic competence as a cognitive ability to

assess and plan required to execute the appropriate form of language use in a successful way and this ability is used in a more conscious way by language learners. In the updated version of the model (Bachman and Palmer, 1996), authors consider strategic competence as a set of metacognitive components or strategies that provide a cognitive management function in language use. The areas of metacognitive strategy use are goal-setting component, assessment component, and planning component. Strategic competence in this updated model includes both compensatory and non-compensatory behaviours. Strategic competence is viewed as a distinct ability to make use of linguistic resources in a specific context to achieve a plan and it also involves general cognitive abilities such as problem-solving abilities (Bachman, 1990). The next section will focus on the description of strategy and distinguish reading and test-taking strategies in a second language from each other.

2.3 Strategies in L2 reading tests

Before making a distinction between reading and test-taking strategies, it is necessary to define what a strategy is and what makes it different from skills that are used during the reading process. While strategies are defined as conscious problem-solving activities, skills are defined as subconscious automatized abilities in reading (Urquhart and Weir, 1998; Cohen, 1998). In another definition, strategies are described as "deliberate, goal-directed attempts to control and modify the reader's efforts to decode text, understand words and construct meanings of text." (Afflerbach, Pearson, Paris, 2008, p. 368). Making use of strategies helps reader examine the strategy, monitor its effectiveness and revise the goals whenever necessary. The authors also underline the fact that using strategies doesn't always make the reader a successful one. However, the skills are automatic actions that

usually occur without awareness of control of components involved. While the main distinction between a skill and strategy seems to be the element of "consciousness" (Cohen and Upton,2006; Phakiti, 2003; Urquhart and Weir, 1998), Grabe and Stoller (2002) point out that strategies can be automatized and that is the reason why "strategies for definitional purposes are best defined as abilities that are potentially open to conscious reflection and use…" (p.15-17). It is necessary to note that the element of "consciousness" will be accepted to distinguish strategy and skill from each other in this study in terms of data analysis. A second distinction that must be noted and that is relevant to this study regarding the strategy use is the one between reading strategies and test-taking strategies. Both types of strategies will be discussed in the following subsections.

2.3.1 Reading Strategies

Reading strategies are those that a reader needs so as to comprehend a text along with processing skills (Koda, 2005). Investigating the use of these reading strategies can illuminate our understanding regarding how a reader interacts with a text and how their choice of strategies affects their comprehension of the text. In this context, the analysis of reading strategies can reveal the extent to which a reader understands the purpose of the reading, how they continue to understand the text and what they do when there is a comprehension breakdown (Cohen and Upton, 2006). Given that understanding the use of reading strategies can reveal valuable information about how a reader copes with the reading task at hand and how the reading task is processed, it is necessary to understand what these reading strategies are. In terms of classifying the reading strategies, there have been diverse attempts.

To start with, Pressley and Afflerbach (1995) divided reading strategies into three broad categories that are planning and identifying strategies, monitoring strategies and evaluating strategies. Planning and identifying strategies are the ones that allow the reader to construct meaning of the text such as planning how to read, checking the goal of reading, checking prior knowledge about the text, etc. Monitoring strategies are monitoring ongoing understanding of the text, predicting, rereading, etc. As to the evaluating strategies, they are used by the readers to reflect or respond to the text. Chamot and O'Malley (1994) categorized reading strategies as cognitive, metacognitive and social affective strategies. Purpura (1999) investigated the relationship between perceived metacognitive and cognitive strategy use in general contexts and language test performance based on human-information processing theory by using structural equation modelling approach. The results of the strategy use questionnaires that were applied to 1.382 test takers before taking the tests showed that cognitive strategy use was a multidimensional construct consisting of comprehending strategies (strategies to understand the text such as identifying main ideas, translating, making inferences), memory (storing information such as rereading, note-taking, paraphrasing) and retrieval strategies (related to recalling information such as using prior knowledge, applying grammar rules) while metacognitive strategy use was a unidimensional construct that is comprised of planning strategies (setting goals, directing goals, planning beforehand, etc.), monitoring strategies (checking comprehension, noticing comprehension failure) and evaluating strategies (assessing and evaluating actions and performance in reading comprehension). It was found out that metacognitive strategy has a direct influence on cognitive strategies and the effect of cognitive strategies is mediated by metacognitive strategies. McNamara, Ozuru, Best, O'Reilly (2007) suggest a 4-

pronged comprehension strategy framework based on the Construction-Integration model by Kinstch (1998) and monitoring comprehension and reading strategies lies in the centre of this framework. Four categories of strategies that constitute the prongs of the framework are preparing to read, interpreting the words, sentences and ideas in the text, going beyond the text and lastly organizing, restructuring and synthesizing information in the text. Preparing to read is about setting or recognizing the goals of reading and using pre-reading strategies. The category of interpreting the words, sentences and ideas in the text refers to creating a text-based level of understanding. Going beyond the text means connecting the text with prior knowledge and the last category includes strategies that help reader organize, restructure and synthesize the information in the text. McNamara et al. (2007) emphasize that this framework is based on the idea that "reading strategy use is intrinsically metacognitive with the monitoring of the comprehension at its core. Metacognitive reading strategies induce and support the use of monitoring which in turn facilitates the use of various reading strategies" (p.467). Therefore, this framework also confirms the relationship between cognitive and metacognitive reading strategies and how they are interrelated with each other.

Phakiti (2003) investigated the relationship between the use of cognitive and metacognitive strategy use and reading test performance. 384 Thai students completed the questionnaire after they completed the exam to find out their use of strategies during the exam and it was found out that there was a positive correlation between cognitive and metacognitive strategy use. Phakiti (2008) examined the hierarchical relationship of strategic competence as strategic knowledge of cognitive and metacognitive strategy use in general (trait) and strategic use in an actual language use situation (state) using a fourth-order factor model of strategic

competence using structural equation model. A strategy questionnaire prepared in simple present tense was administered before the exam to investigate their general strategy use (state) and at the end of the test, a strategy questionnaire prepared in simple past tense was administered to explore the strategies used at that specific exam to 561 Thai students. It was found out that strategic competence affects trait metacognitive strategy use which regulates trait cognitive strategy use and state metacognitive strategy use. State metacognitive strategy use affects state cognitive strategy use. The results demonstrated that strategic competence is multi-faceted and highly complex.

At this point, it is important to mention that in Cohen and Upton's (2007) categorization of strategies, reading strategies (language learner strategies) are categorized as one of the test-taking strategies. However, in this study, there is a distinction between reading and test-taking strategies. Reading strategies are accepted as cognitive and metacognitive strategies as explained above and test-taking strategies will be discussed in the next subsection.

2.3.2 Test-taking strategies

Test-taking strategies are described as those "consciously-selected processes that the respondents used for dealing with both language issues and the item-response demands in the test-taking tasks at hand" (Cohen, 2012a, p.1). Cohen further asserts that three types of strategies, which are language learner strategies, test-management strategies and test-wiseness strategies, come into play when responding to language test items (Cohen, 2013). He mentions that the first type of strategies cannot be called as test-taking strategies at all as this category consists of language learner strategies and these strategies are considered to be a part of one's language ability.

These are viewed as similar to compensatory strategies discussed in Chapter 2.2. The other two strategies are test-management strategies and test-wiseness strategies. Test-management strategies are used for responding meaningfully to test items and tasks and hence, they include consciously selected processes to produce the correct answer. Some of these strategies might include going back and forth between the passage and the question to find more information about what to look for or dealing with multiple-choice items systematically to consider all the distractors and to devise a rationale to explain why one is a better choice than the others (Cohen, 2012b; Cohen, 2013). Test-wiseness strategies include the use of testing formats or other peripheral information to come up with an answer. Some of these strategies that are relevant to multiple-choice tests are stem-option cues (when matching between the stem and an option is possible), selecting an option just because it has a key word from the passage in it, selecting an option out of a vague sense that other options cannot be true, similar options in the distractors (eliminating distractors as they mainly say the same thing) (Allan, 1992). Cohen (2013) remarks that test-management strategies contribute to construct-relevant (desirable) variance whereas the purpose of test-wiseness strategies is to help test takers respond to items and tasks without utilizing the competence in the targeted language skill or without engaging in the intended processes required by the item. That is why, test-wiseness strategies are considered as resulting in construct-irrelevant variance. Multiple-choice items are more likely to be facilitated by the use of test-wiseness strategies and for this reason, this type of item needs to be piloted and revised carefully to prevent any misinterpretation of the scores. The study by Yang (2000) proves this to be the case. Three hundred and ninety Chinese students responded to a modified version of Rogers and Bateson's (1991, in Yang 2000) test of test-wiseness (TTW)

and a TOEFL test and as a result of the tests and verbal reports, it was seen that the majority of items were susceptible to test-wiseness. The students who were labelled as test-wise students had a more meaningful, thoughtful, logical and less random approach to the items and they were more sensitive to subtle clues of test-wiseness. In another study, 43 students in Taiwan (Tian, 2000) produced verbal reports while they were doing the test tasks. It was found out that low-scorers depended on mostly word-level strategies while high scorers focused on the understanding of the reading passage and used test-taking strategies as an auxiliary to support themselves in the response process. The author concluded that comprehension and metacognitive strategies should be the focus of instruction. General test-taking strategies should be taught and other strategies should be encouraged only as an aid to general comprehension strategies. Therefore, it is claimed that clever respondents can use test-wiseness strategies in moderation to answer the test items correctly without going through the required processes by test items or functioning in the language (Cohen, 2013).

Any discussion on test-taking strategies without mentioning the test format effect would be incomplete as it is known that the use of test-taking strategies is evidently related to the test format effect (Sarnaki, 1979). To this end, next section will discuss test-format effect under the light of the empirical studies.

2.4 Test-format effect

Evaluating L2 reading ability is seen as problematic due to the fact that comprehension cannot be observed directly or at the extreme, it is claimed that "comprehension cannot be measured" (Smith, 1994, p.53). Readers must perform something as an indicator of their comprehension (Johnston, 1984) and in addition to

this, the number of testing formats is limited. Cohen (1998) classifies test formats as more indirect and more direct testing formats. He asserts that indirect testing formats are described as the ones that don't reflect real-world tasks and that is the reason why they can prompt the use of test-taking strategies by which test takers try to cope with the demands of test format. Multiple-choice questions or cloze tests can be seen as an example of this indirect testing format. In addition, it is claimed that in such formats, to come up with correct answers, readers can make use of extra mental processes or item-responding processes along with the genuine reading processes that the item intends to elicit (Lim, 2014). Test-wiseness strategies can be given as an example for such processes. In open-ended questions or summary tasks, which are seen as more direct testing formats, the test taker is required to produce an answer based on the comprehension of the text. However, Cohen (1998) suggests that in this type of question, test takers are prone to copy the material directly from the text as a response and this evokes the question of whether the test taker really understands the material or it is just some kind of surface matching. Despite such concerns, Rauch and Hartig (2010) assert that open-ended responses can reflect the classroom reality better as they can mirror teacher and student communication on text more closely.

There are a great number of studies that have contributed to the discussion regarding test format effect. Martinez (1999) asserted that test item formats are different in the range of cognitive demand and in the range of cognitive demands they require. It is asserted that multiple-choice format elicits low-level cognitive processing while more complex thinking is required by open-ended formats. Paulson and Henry (2002) further suggested that completion of cloze items alters the normal reading processes of test takers who want to complete the test successfully as it can be suggested by evidence from eye-movement measures. Rodriquez (2003)

conducted a meta-analysis based on the previous studies that compared the trait or construct-equivalence of multiple-choice and open-ended questions. There were 29 studies including 56 correlations between items in both formats. The results indicated that in the items constructed in MC or OE format with the same item stem (stem-equivalent), there was a high correlation between two formats. The second highest correlation was found between content equivalent designs between the two test item formats. The lowest correlation was found between non-content equivalent designs. Authors concluded that different scores for each item format should be reported in the studies when both items are not developed to be stem-equivalent. Rupp, Ferne, Choi (2006) claimed as a result of the cognitive interviews they conducted with the test takers that multiple-choice format elicits response processes from test takers that are strikingly different from the ones that respondents would draw on while reading in non-reading situations. Multiple-choice tasks were seen as a process of problem-solving rather than reflecting comprehension on the grounds that readers made use of unconditional and conditional strategies on purpose to cope with the task demands and to determine the correct choice. MC format also led to a more segmented reading. A meta-analysis by In'nami and Koizumi (2009) investigated the effects of open-ended and multiple-choice question formats on L1 reading, L2 reading and L2 listening comprehension. Based on the results of 56 data sourced, it was concluded that multiple-choice formats are easier than open-ended formats in L1 reading and L2 listening. However, there were not any format-effects on L2 reading. Rauch and Hartig (2010) analysed the dimensionality of a reading comprehension test with non-stem equivalent multiple-choice and open-ended formats. The results of their study indicated that multiple-choice and open-ended formats differ in their cognitive demands because multiple-choice format and open-ended format loaded on a general

latent dimension which measured abilities necessary to master basic reading processes. However, open-ended format was correlated with nested latent dimension which measured abilities necessary to master higher reading processes unlike multiple-choice items. Contributing to the discussion between these two item formats, Field (2011) suggested that test-taking strategies that are specific to multiple-choice format can undermine the cognitive validity in a test. In another study by Ozuru, Briner, Kurby, McNamara (2013), authors concluded that multiple-choice format and open-ended format measure different aspects of comprehension process. In this study, test takers read a short text while explaining preselected sentences and then they answered multiple-choice and open-ended questions. The results showed that performance on open-ended formats correlated with self-explanations while multiple-choice performance was in correlation with topic-specific prior knowledge from the text and there was little relationship between performance on open-ended and multiple-choice questions. Lastly, regarding the controversy about open-ended and multiple-choice formats, Prince (2014) asserted that open-ended formats should be favoured more in terms of test authenticity as they can imitate real world tasks better than multiple-choice tasks and this claim was also made by Rauch and Hartig (2010) who asserted that open-ended formats reflect classroom reality in terms of teacher and student communication.

In this context, it is necessary to keep in mind that different item formats due to the difference in their nature can elicit extra mental processes from the test taker and therefore, they may tap into different aspects of reading construct. As a result, tests may even involve construct-irrelevant elements by making the readers engage in the mental processes not intended by the test developers. Therefore, understanding the processes that readers go through or different kinds of strategies they utilize

while responding to these two different item formats, namely open-ended and multiple-choice questions, is of utmost importance in terms of test validation. There are basically two methods of investigation that are widely used and have been informative in understanding the underlying cognitive processes reading items trigger; verbal reports and eye-tracking methodology. Significant literature on both of these techniques will be reviewed below.

2.5 The use of verbal reports

Verbal reports have been used extensively in the first (L1) and second (L2) language research to provide a deeper insight into issues such as language learners' cognitive processing, thought processes, and strategies for a long time. Psychologists Ericsson and Simon (1993, p. xi) assert that "both concurrent and retrospective verbal reports are now generally recognized as major sources of data on subjects' cognitive processes in specific tasks". Bowles (2010) simply defines verbal reports as a learner's comments either while completing the task or some time after the task is completed. Verbal reports in which verbalizations and task completion occur simultaneously are called concurrent reports and the ones which are collected after the task completion are called retrospective reports. As to another method of collecting verbal data that is called stimulated recalls, Gass and Mackey (2000) describe them as a subset of retrospective reports that occur after the completion of a task which includes a video- or audio-recording as a stimulus. Both verbal reports and stimulated recalls are seen as introspective methods in which data can be elicited about thought processes in completing a task or activity. The assumption underlying this methodology is that humans have access to their internal thought processes and

they can verbalize these processes (Gass and Mackey, 2000). In addition, Cohen (1998, p.34) classifies verbal report into three types that are as follows:

(1) *self-report:* learners' descriptions of what they do, characterized by generalized statements about learning behaviour – e.g. "I tend to be a speed listener,"

(2) *self-observation:* the inspection of specific, not generalized, language behaviour, either introspectively, i.e. within 20 seconds of the mental event, or retrospectively – e.g. 'What I just did was to skim through the incoming oral text as I listened, picking out key words and phrases', and

(3) *self- revelation:* 'think-aloud' stream of consciousness disclosure of thought processes while the information is being attended to – e.g. 'Who does the "they" refer to here?'

Green (1998) categorizes verbal protocol analysis or verbal reports in terms of form of report – i.e. think-aloud (including information in both verbal form and non-verbal form such as the spatial location of an item in the text) or talk-aloud (only words in the mind or thoughts), temporal variations – i.e. concurrent (data collected at the same time with the task completion) or retrospective (after task completion) and finally in terms of procedural variations such as being mediated (the use of probing questions about the task) and non-mediated (non-intrusive prompts as much as possible such as requests "keep talking"). In this context, think-aloud is regarded as being more advantageous than talk-aloud data as it includes non-verbal data, too. As to the timing of the verbal report, while Green (1998) suggests the use of concurrent verbal report instead of retrospective verbal data as the latter may include more intervening variables. When there is a delay between task completion and production

of verbal report, retrieval process can be considered as being “fallible” and also participants may filter information to tidy it up for a retrospective report, report what they believe the researcher requires, omitting or forgetting important information, or adding extra information or processes to form a complete report (p.6). However, reporting especially – on-line reporting (also called self-revelation) may alter the original thought process more than when there is no recoding (Faerch and Kasper, 1987, p.19 as cited in Cohen, 1998). Similarly, concurrent verbal report may also result in “reactive effects” by producing data no longer reflecting the intended processes in a task because it may distort the process of reading as it makes readers read more closely than normal by concentrating on the additional cognitive and metacognitive task (Mann, 1982 as cited in Cohen, 1998). Bowles (2010) suggests that the threats of collecting retrospective verbal reports can be minimized if there is “a short delay between task performance and verbalization” (p.14). Lastly, Green (1998) asserts that questions asked to probe or mediate can switch individual attention to what it is required from them. It is further suggested that even if probes are to be used, they need to be carefully worded to reduce the likelihood of intervention which can change the process of natural sequence of information. Cohen (1998) also adds that instructions should be carefully chosen to ensure that particular cognitive behaviours are elicited and participants should also be given warm-up trials until they are not confounded with explanations or justifications. Gass and Mackey (2000) remark that participants may be given simple instructions and a direct model during the training or warm-up phase. Cohen and Upton (2006) similarly indicate that it is necessary to brief participants on what is required of them and inform them about the procedure to be used. While Green (1998) supports that there must be minimum intervention from the researcher during verbal reports, Pressley and

Afflerbach (1995) assert that prompting respondents to use particular processes might be necessary at times. As to the language of verbal reports, Bowles (2010) points out that the language of verbal report should be in second language only if it is the only language between researcher and participant and if it is required by the research question. Cohen (1998) also supports this argument by asserting that if the participants are speakers of different languages or obtaining translations is unfeasible, target language can be used in verbal reports. However, he cautions that researchers collecting verbal reports in the target language may be at the expense of collecting sufficient data.

Despite all the controversy about the use of verbal reports to investigate the cognitive processes as to the timing, language or probable effects of it on the task completion, Cohen (1998) concludes by referring to previous studies making use of verbal reports:

> Whereas the reliability of mentalistic measures has been questioned in comparison with behaviouristic measures, research has demonstrated that verbal reports, elicited with care and interpreted with full understanding of the circumstances under which they were obtained, are, in fact, a valuable and a thoroughly reliable source of information about cognitive processes. (p.38-39)

A deeper and comprehensive insight into these cognitive processes can be useful in many different ways such as phases of test development, evaluation or justification of using certain item types, identifying item and task characteristics, construct validation process (Green, 1998). Finally, it is important to note that verbal report is not seen as a replacement for other means of research but it might also be used as a complementation to them as all research measures have certain types of strengths and weaknesses (Green, 1998; Cohen 1998). Two studies which made use of verbal reports will be reviewed in detail below.

Rupp, Ferne and Choi (2006) set out to explore the possible effects of assessing reading comprehension with multiple choice format. Cognitive interviews were conducted with 10 participants (3 males and 7 females) from second language courses at a large Canadian University in Ontario. Semi-structured interviews used as prompts were administered to the participants to collect verbal reports from them while they were responding to reading comprehension tasks. The tasks were chosen from *CanTEST,* a large-scale test developed in Canada. The participants completed three tasks, one of them was accepted as a preparation test while in the other two texts, the participants were first asked to read the questions or the text and then they answered the questions. Researchers asked them how they answered the questions one by one and asked them to rate the difficulty of each item. Interviews were digitally recorded, transcribed and read into NVivo software. The transcripts were analysed according to the strategies used in the questions. The results showed that strategy selection was divided into two categories such as micro-level strategies (utilized according to perceived characteristics of the text and questions; text was a more influential factor) and macro-level strategies which can be summarized as unconditional and conditional strategies. In both conditional and unconditional strategies, the process of answering heavily depended on key-word matching. Authors concluded that responding to multiple-choice questions was never a linear process in which test takers first read a text to form a text-base and then answered questions (focusing on microstructure rather than macrostructure) and texts were scanned only for key words and therefore, it was suggested that segmentation and localization were main functions of many types of multiple-choice questions. Reasoning process evoked by multiple-choice was unique to a testing process.

Cohen and Upton (2006) aims to describe reading and test-taking strategies that are used by test takers to complete the reading tasks in the reading sections of LanguEdge course materials. The study mainly tries to describe the use of strategies while answering single-selection multiple-choice format used for basic comprehension and inferencing questions and the new selected response reading to learn items in TOEFLiBT. The participants consisted of 32 international undergraduate and graduate students. They first completed a background questionnaire and then they took LanguEdge training and LanguEdge pretest to determine proficiency. Secondly, they got training on giving verbal reports and then they completed real tasks and gave concurrent and immediate verbal reports. The authors argued according to the results that all types of questions, namely basic comprehension questions, inferencing and reading-to-learn questions assessed the same academic skills, not different ones as it was expected, through a close analysis of strategy use. Results also showed that test takers used problem-solving approaches to answer the questions and they didn't want to learn anything from the reading texts. On the other hand, the claim made by ETS as to the TOEFLiBT was confirmed as the questions required both local and global comprehension of the text. For basic comprehension vocabulary questions, it was emphasized that test takers mostly relied on their background knowledge. In addition, test takers mostly used test-management strategies compared to reading strategies or test-wiseness strategies. The limited use of test-wiseness strategies was thought to strengthen the validity of TOEFLiBT. As to the reading-to-learn questions, it was observed that these questions were regarded as less difficult, which contradicts with the intentions of the test developers. The primary reason for this was thought to be the fact that readers become very familiar with the test content until they solve this type of questions.

2.6 Eye-tracking technology

The use of eye-tracking technology seems to be quite helpful in elucidating reading processes. Regarding the use of this technology, Staub and Rayner (2007) assert that eye movement data is viewed as quite informative to understand reading. The authors further assert that eye movements can "provide a moment-to-moment indicator of the ease (or the difficulty) with which readers are able to comprehend the text that they read." (Staub and Rayner, 2007; p.327). In this respect, Spivey, Richardson and Dale (2009) argue that eye-movements can be considered as good indicators of cognitive processes and eye-movements are seen as a window into language and cognition. Solheim and Uppstad similarly assert that:

> Eye-movement recordings of reading on a discourse level yield an on-line record of the reading process in the form of information about what readers visually focus on in the text passage and for how long they inspect different passages. In an assessment situation eye-tracking data can provide on-line information about readers' decisions to search the text in order to give an answer to a question, and about how accurate and effective that search is. (Solheim and Uppstad, 2011, p.155)

In the quotation above, it can be obviously understood that eye-tracking data can yield valuable information regarding the cognitive processes of test takers as there is minimal disturbance to test takers' cognitive processes and they are allowed to complete the tasks as they would do under non-testing conditions. However, it is also suggested by some researchers that eye-tracking data should be accepted as an indicative of cognitive processing, rather than a full reflection of it (Reichle, Warren, McConnell, 2009).

It is evident that the use of eye-tracking seems as a promising way of investigating the processes in reading tests and helpful in validation research. It is assumed that eye movements are related to cognition (Rayner, Reichle, Pollatsek, 2005; Reichle, Rayner, Pollatsek, 2003). Rayner (1998) suggest that there is close

link between the point in a text where our eyes fixate and the focus of our attention. Concordantly, eye-mind hypothesis suggests that there is a strong association between the eye movements and mind (Just and Carpenter, 1980) and E-Z model points out that gaze position properly indicates the linguistic features to which readers attend (Rayner, 2009). Therefore, it is worth understanding how eye-tracking is done and what can be investigated through this technology.

Eye movements are recorded by measuring the movement on the cornea and pupil. Infrared light is reflected via a mirror into one of the participant's eyes and in this way, this creates a reflection of the retina and cornea. The corneal and retinal reflections are used to calculate where the participant's eye is focused. When people read something, their eyes are not gliding smoothly but in fact they make a series of jumps and they remain stationary between these jumps. These rapid movements or jumps are known as saccades. The saccades require 20-40 milliseconds or 7-9 letter spaces in normal English on the average saccade (Staub and Rayner, 2007; Rayner, Chace, Slattery, Ashby, 2006). It is further stated that the duration of the stationary periods, which are called "fixations" comprises a somewhat right-skewed normal distribution with the minimum at about 50-100 milliseconds and maximum about 500 milliseconds, with the mean at around 200-250 milliseconds. It also vital to note that meaningful information is extracted from the text during the fixations, but during saccades, the visual system does not register the information picked up by the retina (Rayner,1998; Staub and Rayner, 2007). There is great variability in saccade size ranging from 20 characters to only a single character. In skilled readers, about 90 % of saccades move forward, with the rest moving the eyes backward either to solve comprehension difficulties or to correct error in the programming of forward saccades (ibid.). These backward movements for a distance of a few letters to

reprocess a word are called regressions. Regressions of more than a few letters are indicative of reader's probable failure to understand the content. Another backward movement called "return sweeps" are defined as eye's return to a precise fixation point recalled by the reader as causing difficulty. Also, readers backtrack through the text until they find what causes difficulty. Lastly, the corrective saccades are eye movements which tend to re-identify the text. These four types of backward movement are used to correct insufficient reading (Rayner, 1998). As a text becomes more difficult to understand, fixations become longer, saccades shorter and regressions occur more frequently (Rayner, 1998). If there are multiple fixations and long durations, it implies that there are comprehension difficulties (Paulson and Henry, 2002 as cited in Lim, 2014). Two experiments conducted by Rayner, Chace, Slattery, Ashby (2006) confirm the claims that overall text difficulty affects eye movements such as average fixation duration and the number of fixations and eye movement measures are sensitive to global passage difficulty (experiment 1) and larger regressions reflect comprehension failures to a great extent, not short regressions as shown by the example of semantic illusions (experiment 2).

The number of fixations on a word depends on the word length (Brysbaert and Vitu, 1998; Rayner and McConkie, 1976) and very short words and function words can be skipped (Carpenter and Just, 1983). With regard to the fixations, a reader of English is likely to get useful word identity information mainly about the word that is being fixated and the word that is just right to this fixation. The word just right to fixated point is processed only in terms of the specific letters, word's meaning or its morphological composition is understood only through fixation (Staub and Rayner, 2007). The text visible on each fixation can be divided into three regions. The foveal region extends to 1° of visual angle to the right and left and this

area covers 3-4 letters to the left and right to the fixation where the acuity is the sharpest. Beyond this area, acuity drops but readers can still obtain letter identity information in parafoveal region. This area extends 5º of visual angle to the right and left of the fixation point. Beyond this area, readers are only aware of the shape of the text, such as where a line ends (Rayner, 1997; Staub and Rayner, 2007). It is also asserted that readers don't make use of the information from the lines of the text below the one they are reading (Pollatsek, Raney, LaGasse, Rayner, 1993).

Another point that is worth mentioning is that eye movements in reading can be affected by the properties of individual words such as frequency, familiarity, age of acquisition, meaning ambiguity, morphological decomposition and predictability (Rayner, 1997; Staub and Rayner, 2007). For instance, readers look longer at the words that are relatively infrequent while they skip high-frequency words more often. Other factors that affect eye movements are syntactic processing and discourse processing. With regard to the syntactic ambiguity, the results are inconclusive in that while it leads to a slowdown in some cases, it may also lead to fast reading times in some other cases (Staub and Rayner, 2007). As to the discourse processing, it involves determining what pronouns and definite descriptions refer to or making inferences about relationship between events, explanatory, causal and chronological relationship (ibid.).

With regard to the eye-movement measures, the common measures for word reading are first fixation duration (the duration of first fixation on a word), single fixation duration (those cases when only a fixation is made on a word), gaze duration (the sum of all fixations before moving to another word) and total fixation time (the sum of all fixations including regressions) (Rayner, Pollatsek, Ashby, Clifton, 2012; Rayner, Chace, Slattery, Ashby, 2006). When experimental questions focus on

sentence or discourse processing and unit of analysis is larger than a word, first pass reading time (the sum of all fixations in a region) and total reading time (the sum of all fixations in the region) are typically computed. In addition, go-past time, which is the sum of all fixations from first entering an area until exiting in forward direction, can be calculated (Rayner, Pollatsek, Ashby, Clifton, 2012; Rayner, Chace, Slattery, Ashby, 2006). Along with these measures, Jarodzka and Brand-Gruwel (2017) suggest that researchers also investigate "specific words or parts of the text that a participant did or did not read carefully, which entail some kind of crucial information by means of AoI analyses" (AoI: Areas of interest) (p.195). Besides, "reading depth" can be looked into as another important measure in which how much of a text is read and how much of it is skipped (Holmqvist et al., 2011 as cited in Jarodzka and Brand-Gruwel, 2017).

The studies which made use of eye-tracking technology so as to investigate test format effect or validate a test will be reviewed. Paulson and Henry (2002) attempted to investigate the reading process that test takers go through as they complete a reading assessment called DRP (Degrees of Reading Assessment) which consists of cloze items measuring reading comprehension through the use of eye-tracking technology. To this end, the study examines the eye movements of 10 participants who were undergraduate students at a large university in the United States (mean age 19). Then, they read two baseline passages from DRP exam without any questions (the answers of the cloze items were replaced in these two texts). In the last text, there were cloze blanks and choices for each blank as it was shown in the test itself. The results showed that the eye movements of students taking the DRP don't correspond to their unclosed baseline reading in any way. DRP modified cloze process itself demands an interruption in the normal flow of reading by demanding

regressions and the authors interpret these backward movements as the use of test-taking strategies rather than the use of reading strategies and they claim that this is not consistent with normal reading comprehension process. They conclude that this test changes test takers' reading processes to complete the items successfully.

Solheim and Uppstad (2011) aimed to explore what parts of text participants read to answer questions that required the integration of the text and visual stimuli through the use of eye-tracking technology. Thirty-four students who were in the seventh grade (mean age 12.75 years old) participated in the study. The text was taken from a science book and to understand the text, the participants needed to gather both verbally and pictorially presented information from various parts of the page and integrate them into a comprehensive whole. The items consisted of both constructed-response items and multiple-choice items. Gaze durations were analysed. The results showed that there was no significant relationship between how much time readers spent reading the text or answering the questions and how well they scored in the test. Secondly, there were no significant differences in terms of the use of integrative saccades (this type of saccade refers to the transition between verbal parts of the text and illustrations) contrary to the expectations. First-time readers and non-strategic readers went through the same phases while answering the questions but they got different results. Both types of the readers didn't read the relevant parts of the text but first-time readers got it correct while non-strategic readers missed the question. In a similar vein, both effortful readers and task-oriented readers read the relevant parts of the text and while task-oriented readers did the relevant question correctly while effortful readers missed that question. Authors tried to form a comprehensive picture of how test takers tried to answer the questions and the processes they underwent using the data from eye movements of participants.

The study by Bax and Weir (2012) attempts to find out the cognitive processes employed by test takers on a computer-based CAE reading test to investigate the cognitive validity of the test. By combining eye-tracking data and questionnaire data, the authors tried to understand whether CAE test items could elicit a wide range of appropriate cognitive processes or not based on the model developed by Khalifa and Weir (2009). The retrospective questionnaire and the analysis of eye-tracking data was based on this model. One hundred and three multinational first and second year students from the undergraduate foundation year took part in the study; however, the authors only analysed six proficient learners' eye movement data on five valid items. The eye-movement data were analysed by nine predetermined criteria based on the model by Khalifa and Weir. The criteria are classified as follows (p.8): (a) whether participants read the questions, (b) whether participants read the questions before reading the text (at least three fixations), (c) whether participants used expeditious search strategies to locate the correct place of the answer efficiently, (d) whether participants read all the question options, (e) whether participants read all the options carefully (3 fixations per option), (f) whether participants skim options (fewer than 3 fixations), (g) whether participants fixated or focused most heavily on target, (h) whether participants read more than one paragraph carefully, (i) whether participants scrolled or/and sampled various parts of the text. As a result of the analysis of retrospective questionnaires, the responses were analysed in relation to the eye-tracking data and marked as accurate or inaccurate. Participants were accurate in their self -reports in 68.4% of cases and inaccurate in 31.6% of cases. This proved the limitations of retrospective questionnaires in providing accurate data. The data showed that the test elicited processes at lower-level, sentence paragraph, across paragraphs and whole-text level.

Also, the study emphasized that the use of eye-tracking was very promising in elucidating cognitive processes.

Bax (2013) investigated test takers' cognitive processes while completing an onscreen IELTS reading test items through the use of eye-tracking and stimulated recall interviews. The focus of the project was to find out the difference between successful and unsuccessful test takers' cognitive and metacognitive processing. The study attempted to assess the cognitive validity of the test. It was assumed that successful test takers would use relevant processes whereas unsuccessful readers wouldn't do so. It also investigated the usefulness of eye-tracking technology in revealing cognitive processes of test takers. Seventy-One Malaysian undergraduate students at a UK university participated in the study. The study focused specifically on the analysis of careful and expeditious local reading in terms of Khalifa and Weir's (2009) model. The results of 38 students were randomly chosen for eye-tracking analysis and 20 of the participants who took part in eye-tracking sessions were randomly chosen for stimulated recall interviews. The item analysis consisted of three steps. In the first step, item analysis was run and one item was excluded from eye-tracking data. Quantitative analyses were done comparing successful and unsuccessful readers in terms of total fixation duration, fixation count, total visit duration, visit counts for target items using Mann-Whitney U tests. There were significant differences between these two groups in terms of these four criteria. The results showed that unsuccessful students couldn't read expeditiously to locate information as they spent more time looking for information while successful students showed greater success in locating correct paragraph and focusing on key element. Unsuccessful readers sometimes read faster but couldn't find key

information. The research was also successful in demonstrating the potential of eye-tracking as to the test validation.

Lim (2014) tried to find out to what extent TOEFLiBT measures reading construct and to what extent test format affects students' use of test-taking strategies and how the use of test taking strategies affects reading process and whether these are relevant to the construct. For this purpose, reading componential analysis was implemented to see the extent to which test scores can be determined by sub-reading skills. Eye movement data were analysed to find out cognitive processes. The effect of test format on test performance and test taking processes was investigated. Two-hundred and six Chinese ESL students completed an online survey on the reading difficulties encountered in the academic setting. Ninety participants were invited to eye-tracking data collection sessions. Two comparable reading texts were chosen from TOEFLiBT test practice volume and two multiple-choice testlets were chosen and open-ended versions were created for them. The question stems were kept as equal as possible except negative factual information questions. The sentence simplification questions and insert text questions were excluded as they were not appropriate to transform into open-ended format. In addition, grammar test, vocabulary test, lexical processing task, sentence processing task, working memory test and strategy questionnaires were created and given to participants. Also, stimulated recall interviews and reading-difficulty questionnaires were administered. The measures used for eye-tracking were time to first fixation, fixations before, first fixation duration, total fixation duration, total visit duration, visit count and also the predetermined criteria used in Bax and Weir (2012) to identify reading types. The results showed that reading comprehension was correlated with vocabulary knowledge, grammar knowledge and word recognition skills and the second factor

was test-taking strategy. Three factors that caused difficulty were unknown vocabulary, a lack of background knowledge and allocated time for reading. Almost all readers read the question before the text. Slightly more than half of the students read a text carefully. Search reading was rarely observed and high scorers read the question options but in a strategic manner and they also put their focus on the target areas. Readers rarely read more than a paragraph and expeditious reading at global level was scarce. As a result, careful reading at global level was the dominant one and careful reading skills were used more than expeditious reading skills. Regarding the test format, it was found out that multiple-choice questions were easier than open-ended questions. Total fixation duration on question stem, first paragraph, key sentence and key phrase were longer in the open-ended format than multiple-choice as a result of the comparison between test takers who got full score in both formats. Therefore, test takers paid more attention to the key information in open-ended format. Also, vocabulary items were observed to be solved without looking at the text but rather with the knowledge of collocations or adjacent words in multiple-choice questions. Therefore, these vocabulary items didn't measure inference abilities as intended, but rather vocabulary size. Good readers depended on their stored vocabulary while unsuccessful readers spent more time trying to guess the meaning. Therefore, the validity of these vocabulary items was questionable. As to the test method effect, the researcher concludes that tapping into true reading ability is problematic due to indirect testing methods as these items required extra mental processes.

The last study to be reviewed is by McCray and Brunfaut (2018) who set out to investigate test takers' cognitive processing while responding to banked gap-fill tasks designed to measure the reading ability during online completion of the task

using eye tracker. The study examines types of processing undertaken by high and low achievers to describe the construct better. The reading model to be used was the one developed by Khalifa and Weir (2009). The seven hypotheses that were drawn by the authors based on the reading model mentioned before were tested through a correlational study in which eye-tracking data were used to create processing measures which were used as independent variables and overall scores on banked gap-fill tasks were accepted as dependent variables. This design was used to show the extent of the relationship between type of processing and performance. The participants were chosen on the basis of a stratified design in which the same number of participants from pre-sessional English course, undergraduates and graduates were selected. The data were collected from 28 participants from different backgrounds. In order to collect eye movement and performance data, six banked gap-fill items were used consisting of 24 items in total. The measures used for eye tracking data analysis were (1) mean time fixating on task, (2) mean proportion of time fixating on test, (3) mean proportion of time fixating on sentences containing a gap, (4) mean proportion of time fixating on words surrounding gap, (5) mean proportion of time fixating on word bank, (6) mean number of visits to word bank, (7) gradient of total fixation duration on word against BNC frequency. For the analysis measures 2-7, areas of interest were determined. To find out correlations, Pearson product-moment correlations were used. The results indicated that lower-performing participants didn’t spend more time on the texts and on the sentences with the gaps as it was expected. However, low-achievers spent more time focusing on three words around the gaps. Better achievers made fewer visits to the bank than low achievers possibly implying fragmented processing by lower performers. The study concluded that there was a difference in the processing of low and high achievers as low achievers made

use of lower-processes while responding to this type of items and the increased attention to the words surrounding the gaps was more representative of localized reading and also lower-level processing. The frequent visits to the word bank by low achievers also proved that reading process was not a linear one. Low achievers also spent more time on less-frequent words and this shows that they were engaged in lower-level processes of reading trying to recognize the words. The conclusion drawn was that there was a major difference between high and low achievers in terms of cognitive processing in the local context of the gap and complexity of the words in the word bank and in such cases, low scorers depend on lower-level cognitive processes to a greater extent.

2.7 Conclusion

Under the light of discussions mentioned and the empirical studies reviewed above, it is evident that investigating the processes triggered by an item type is of utmost importance so as to make sound validity claims about a test. Multiple-choice is an extensively used format and in several studies, multiple-choice and open-ended formats have been compared to each other. However, the validity of multiple-choice format has been widely discussed. While the verbal reports have been used to investigate the use of strategies and reading processes in the studies, eye-tracking technology that has been implemented in recent studies is also promising in terms of revealing strategy use and reading processes. The current study will attempt to investigate test format effect in reading texts by a triangulation of eye-tracking technology and verbal protocol analysis. The study compares two different item types in a systematic way through the use of two different methods unlike the

previous studies regarding this issue. The next chapter will focus on the methodological details of the current study.

CHAPTER 3
METHODOLOGY

3.1 Introduction

The methodology section presents the methods and procedures that are used in the current study. The chapter starts with an explanation about the research questions that are investigated. Participant characteristics and the instruments that are utilized in the study for data collection are explained in detail. The chapter ends with an explanation of how the different types of data are analysed.

3.2 Research questions

The main purpose of this study is to find out the differences between two different test formats in terms of assessing reading comprehension and the study attempts to shed light on what test takers do when confronted with different item formats and whether format difference has an effect on their reading processes and whether it causes test takers to engage in extra mental processes such as the use of different test-management and test-wiseness strategies. Firstly, the study set out to explore whether there are any differences in the performance of test takers when they answer open-ended and multiple-choice questions. Secondly, another goal of the study was to find out how much of the text is carefully processed and what reading processes are utilized by test takers through recording eye movements of respondents. Thirdly, the study aimed to explore differences in reading and test-taking strategy use when respondents answer questions in two different formats through the use of retrospective verbal reports given by test takers immediately after they complete the test. Finally, the study also examined the perceptions of test-takers about both the

two tests (in open-ended and multiple-choice formats) they just completed in terms of the efficiency of their comprehension of the reading texts and the difficulty of the tests. Their opinions were also asked about which item format they thought measured reading comprehension in a better way.

The following research questions were investigated:

1. Does test format have an effect on reading scores?
2. Are there any differences in the percentage of total careful reading time, total reading time (TRT) and total fixation count (TFC) in the whole text across open-ended and multiple-choice formats?
3. Do eye movements (TFC, TRT) of test takers differ in areas of interest that are parts of the text relevant to question (AoI) and question stem (QS) in parallel items in both formats?
4. Which overall reading and test-taking strategies do the test takers report using while answering reading comprehension questions in open-ended and multiple-choice formats?
5. What are the test takers' own perceptions regarding the test format effect in reading comprehension tasks?

3.3 Participants

A total of 34 participants took part in the current study. However, the eye-tracking data of three participants had to be eliminated from the analysis due to calibration problems resulting from various reasons such as participants having eye glasses or slanting eyes and unexpected technical problems. Therefore, the data of 31 participants (27 females and four males) were included in eye-tracking analysis whereas the data obtained from 34 participants (30 females and four males) were

used in verbal reports and semi-structured interviews. The participants consisted of undergraduate students (16 sophomores and 18 juniors) studying at different departments at Boğaziçi University where the medium of instruction is English. The majority of the participants (25 students) were majoring in Foreign Language Education Department in the Faculty of Education, four of the participants came from the Department of Management, two of them were studying Economy and the rest of the students were studying Chemistry, Electrical and Electronical Engineering and Guidance and Psychological Counselling. All participants had normal or corrected-to-normal (through soft contact lenses) vision.

All the participants are native speakers of Turkish and their ages ranged from 18 to 32 (with a mean of 20.8). The majority of them have been exposed to English since they were 10 years old (starting in the fourth grade at primary school). All the participants reported that they were quite familiar with the computers and all the participants (except for one participant) have taken undergraduate courses on computer or technology and it can be claimed that they were computer literate in this sense. In addition, all the participants reported that they had previous onscreen test experience or they participated in eye-tracking experiments before the study. Lastly, all the participants passed Boğaziçi University Proficiency Exam (BUEPT) to become regular students in their departments. It is important to note that a minimum pass mark on BUEPT corresponds to 79 on TOEFLiBT and 6.5 on IELTS academic. All the students participated in the study voluntarily. They were given no course credit if they participated in the study; however, they were given headphones or USB memory sticks as a compensation.

3.4 Instruments

The study will make use of two reading texts adapted from TOEFLiBT practice tests. To collect data, eye tracking methodology, immediate retrospective verbal reports and short semi-structured interviews were administered.

3.4.1 Reading tests

Two reading passages which were accepted as being comparable as a result of the analyses of various textual features were selected from a TOEFL practice book called "The Official Guide to TOEFL® Test Third Edition". The texts were found to be comparable as a result of automatic text analysis tool *Coh-Metrix*, the vocabulary analysis tool *Compleat Lexical Tutor* and readability statistics (Taylor and Weir, 2012; Green, Unaldı and Weir, 2010). The results of text analysis are summarized in Table 2. The genre of both texts was both chosen as expository texts and they were also about similar topics as it is accepted that different genres can have an effect on the cognitive processes. Weir's (2005) validity framework demonstrates that context validity and theory-based validity have a reciprocal relationship with each other.

After two multiple-choice testlets were chosen from TOEFLiBT practice book, an open-ended version was created for each multiple-choice question. All the multiple-choice questions were converted into open-ended format and the items that were parallel in both formats were kept parallel to each other as much as possible in terms of their item type, their question stem (stem equivalence), text span (where the answer is located in the text; how much of the text needs to be read to answer the question which assured content equivalence) and lastly the difficulty level of the items.

Table 2. Text Analysis of Reading Materials Used in the Study

	Text 1	Text 2
Genre	Expository	Expository
Title	The Origins of Cetaceans	Swimming Machines
Word Count	638	631
Readability		
Flesch Reading Ease Score (0-100)	56	60
Flesch-Kincaid Grade Level (0-12)	9.2	8.5
Coh-metrix L2 Readability	11	11
SMOG	8.7	8
Vocabulary Complexity		
K1+K2 Word Percentage	78.16	80.79
AWL Percentage	4.05	4.44
Type and Token Ratio	0.43	0.45
Lexical Density	0.6	0.58
Text Features %		
Narrativity	14	22
Syntactic Simplicity	80.78	68.79
Word Concreteness	86	85
Referential Cohesion	22.36	20
Deep Cohesion	42	55

At this point, the questions which couldn't be transformed into open-ended format such as negative factual information, sentence simplification questions and insert text questions (i.e. cohesion questions) had to be eliminated. Open-ended versions of inference questions were devised by providing the statement that can be inferred from the text and asking test takers to find the sentence that supports the inference in

the best possible way. The reading tests and the items that were created were sent to five English language teachers who have also experience in test development. These teachers were asked to give expert opinion on both these two texts in terms of their comparability and also on the items with regard to their content, explicitness, ease and the location of the answers (how much of the text needed to be processed to find the correct answer) with the help of cognitive and contextual proforma taken and adapted from Wu (2011) (Appendix A and Appendix B). After the items were revised as a result of the feedback, a pilot exam for open-ended exam was conducted on a group of students who were at the end of the preparatory year at a public university in Istanbul and who were classified as B2 level according to CEFR. The piloting of the first text, "*The Origins of Cetaceans*", was administered with three classes, 62 students while the piloting of the second text, "*Swimming Machines*" was conducted on three classes, 55 students. The results of reliability analysis for both tests can be seen in Table 3 and Table 4. The items which lowered reliability were eliminated and their equivalents were also taken out from the multiple-choice format. In Text 1, items 2,3,5 were eliminated whereas in Text 2, items 1,2,3,5 were eliminated. In this way, there were two different texts (Text 1 and Text 2) and each text had open-ended and multiple-choice forms; for Text 1, there were two versions, one with multiple-choice questions and the other one with open-ended questions. Similarly, for Text 2, there were two forms, one with multiple-choice questions and the other one with open-ended questions. In each form, there were six questions in total as a result of revisions. The questions for Text 1 and Text 2 were also parallel in terms of their item types since they were intended to measure similar processes. The item types and their counterparts are summarized in Table 5. The tests are given in Appendix C, Appendix D, Appendix E, Appendix F.

Table 3. Reliability Statistics for Text 1: Origins of Cetaceans

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
I1	7.32	828.215	0.991	0.966
I2	8.39	1057.98	0.708	0.988
I3	7.19	778.585	0.995	0.966
I4	7.84	923.974	0.982	0.972
I5	6.94	724.225	0.996	0.969
I6	7.61	871.848	0.988	0.968
I7	7.03	743.966	0.997	0.967
I8	7.52	849.795	0.99	0.967
I9	7.42	828.215	0.991	0.966

Cronbach's Alpha was 0.974.

Table 4. Reliability Statistics for Text 2: Swimming Machine

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
I1	4.64	5.643	0.235	0.723
I2	3.95	5.756	0.186	0.729
I3	3.98	5.463	0.322	0.713
I4	4.47	4.884	0.522	0.681
I5	4	5.593	0.233	0.725
I6	4.69	5.551	0.371	0.709
I7	4.16	3.843	0.494	0.705
I8	4.71	5.618	0.362	0.711
I9	4.13	4.854	0.538	0.678
I10	4.47	4.55	0.706	0.648

Cronbach's Alpha was 0.725.

Table 5. Items of Text 1 MC & OE and Text 2 MC & OE

TEXT 1 (FORM MC& OE)	ITEM TYPE	TEXT SPAN (Where the answer was found)	TEXT 2 (FORM MC & OE)
Q1	Factual	One sentence (Local)	Q6
Q2	Factual	One paragraph (Global)	Q3
Q3	Vocabulary	One paragraph (Global)	Q4
Q4	Factual	Across sentences (Global)	Q5
Q5	Inference	One paragraph (Global)	Q1
Q6	Factual	Across paragraphs (Global)	Q2

3.4.2 Recording eye movements

The current study used Tobii eye tracker x1 Light. The Tobii x1 Light Eye Tracker has a variable data-rate which is typically between 28 and 32 Hz, i.e. typically between 20 to 32 data samples are collected per second for each eye. A binocular camera was attached to the bottom of monitor to record eye movements. Gaze accuracy under ideal conditions is 0.4^{o} with binoculars and gaze accuracy is 0.06^{o} with a noise reduction filter. In this study, a chin-rest was attached to the eye-tracker so as to make sure that data won't be distorted as there will be two different tests respectively. The minimum fixation duration for the current study was set at 100 ms (Sereno and Rayner, 2003) which means that eye fixation was defined as the maintaining of the visual gaze for at least 100 ms and the area up to 5° of visual angle was accepted to be within the perception at the time of a fixation.

3.4.3 Retrospective verbal reports and semi-structured interviews

After the test is completed and eye movements are recorded, the text and the questions reappear on the screen and by reminding the participants of the answers they have given, the participants are expected to report how they answered each question. The verbal reports were given in Turkish, which is the native language for both the researcher and the participants and they were audio-recorded and transcribed. The purpose of verbal reports was to identify the strategies used while responding to questions and also confirm or supplement the identification of reading processes as much as possible.

As to the semi-structured interviews, upon the completion of verbal reports, the researcher held a short and friendly conversation with the test takers about their own perceptions about the test format both in the specific tests they had just answered and their perceptions about the effect of test format on reading comprehension in general. The aim of this interview was to understand the test takers' own perceptions about test format effect and to gain a deeper insight into their strategy use as much as possible.

3.5 Procedure

Data were collected individually from the participants, so only the researcher was in the lab during the administration of tests and data collection. There were two conditions in total and in each condition, there were 17 participants. Four forms of tasks were counterbalanced in each condition. Half of the participants took Text 1 MC and Text 2 OE and the other half of the participants were given Text 2 MC and Text 1 OE. There was no time restriction and time of each condition ranged from 60 minutes to 90 minutes depending on the individual performance of the participant.

The experiment took place at the Vision Lab of Psychology Department at Boğaziçi University. In either condition, the process of the experiment consisted of the following steps:

1. The participants first completed a short questionnaire about their background information and signed an informed ethical consent form which was prepared in Turkish, their native language (See Appendix G).

2. During data collection, first, participants adjusted their sitting posture with eyes at a distance of 55 cm from the screen and performed a 5-point calibration. If a calibration point was missing or large errors occurred (e.g. large differences between the gaze point calculated by the eye tracker and the actual dot position), then the erroneous points were recalibrated. The aim of this process was to calculate individual eye patterns and saccades in order to determine their individual pattern of gaze and saccade behaviour for the sake of accuracy of the subsequent tracking.

3. After the calibration was over, the researcher briefed shortly on the following process and what was expected of the participants without revealing the main purpose of the study.

4. Next, the researcher told a short sample verbal report to give the participants an idea about how they were going to report.

5. The participants got a training task which was shown on the computer screen with the Tobii eye tracker attached to it. This training task was prepared both to show each aspect of the process they would follow and allow the participants to practise how to give verbal reports. In the training session, the participants first completed a reading task which had only one question in it. Then, immediately after they had finished the reading task, they started to give a verbal report on how they came up with the answer, what they were thinking during task completion and how they

approached the reading task and the questions in general. The instruction on giving verbal reports was the same for all tasks including the training task and was as follows:

> "Now, I would like you to tell me how you read this reading passage and how you answered its questions. While doing this, I want you to tell me what you were thinking while answering each question as much as you can remember. Try to talk about your thoughts in the order they occurred as much as possible (as much as you can remember). Don't explain why you thought in the way you did or how you should have thought and don't answer the questions again. I just want you to tell me what you were thinking while you were answering each item in a detailed way. We can start whenever you are ready."

6. Upon the completion of verbal reports, the participants were given feedback regarding their performance and how they should continue in giving their reports and what was missing and should be included in their reports. After the students were clear about the process, real experiment started.

7. The experiment started with MC format of reading texts in both conditions. The first step was the recording of eye movements and test takers were asked to read and answer the questions and they were notified that any head or body movement might distort the calibration and they would tell their answers aloud so that the answers were both audio-recorded and the researcher noted them down. Another important detail that is worth noting is that students couldn't scroll down on the page or couldn't choose the answer on the screen not to contaminate the eye-tracking data. In all the tests, the text appeared on the left part of the screen and the questions appeared on the right part (See Appendix C, Appendix D, Appendix E, Appendix F)

and the participants verbalized their answers. It should be noted that there was no interaction between the researcher and test taker during this part.

After the participant told that the test was over, the researcher changed the screen and told that the eye-tracking part was over. The audio-recorder was immediately turned on and the participant was asked to give an item-by-item explanation of what they did and how they answered each question without delay and their answers were reminded to them in the same order as they responded. It is important to note that in some participants, the researcher had to use some prompts which would serve as a guidance such as “How did you answer this question?”, “Can you also talk about the distractors?”, “Keep talking, please”, etc. if the participant kept silent for a time. To avoid interrupting the participant’s recall or self-observations, the prompts were given when they paused.

8. After the MC task was over, the participant took a break for five minutes.

9. The second task which included OE questions appeared on the screen and the same procedure in the step 7 was repeated.

10. After the verbal report was over, there was a short semi-structured interview part in which the participants were asked some questions and there was a short conversation between the researcher and the participant. A summary-list of the whole experiment process can be found in Table 6.

Table 6. Procedure of the Study

Condition 1	Condition 2
Questionnaire & Signing ethical consent Forms	Questionnaire & Signing ethical consent Forms
Calibration	Calibration
Briefing	Briefing
Training Task	Training Task
Text 1 MC Eye tracking	Text 2 MC Eye tracking
Text 1 MC Retrospective verbal report	Text 2 MC Retrospective verbal report
5 minutes break	5 minutes break
Text 2 OE Eye tracking	Text 1 OE Eye tracking
Text 2 OE Retrospective verbal report	Text 1 OE Retrospective verbal report

3.6 Data analysis

First of all, to answer the first research question, a test of ANOVA was conducted to find out whether the differences in the test scores resulted from differences in the method or the text. In this way, it would be possible to understand whether the variance in the scores can be explained by test method effect or not.

Regarding the second research question, to find out how much of the texts is read carefully in each format, careful reading is first operationalized as minimum three fixations on a sentence based on the predetermined criteria by Bax and Weir (2012) to analyse eye-tracking data. To this end, the number of sentences was calculated and it was seen that Text 1 had 40 sentences while Text 2 had 42 sentences in total. The sum of the sentences with at least three fixations was calculated. To determine how much of the text is dedicated to careful reading, the number of sentences that are carefully read is divided by the total number of sentences in the text. The percentages of careful reading in MC and OE formats were compared by Mann-Whitney U test (this test is chosen as normality and homogeneity of variance assumptions are not met) to check whether there is a difference in both

methods in terms of how much of the text is carefully read. In addition, text based total reading time (expressed in seconds) and total fixation count are compared between different methods to see whether there is a difference between them based on test format.

In order to determine the reading processes for each item in both formats which is investigated by the third research question, the results of eye movements will be analysed. In order to decide on the reading processes, the measures to be used in this study are total fixation count (TFC) (frequency count of all individual fixations within a given area of interest) and total reading time (TRT) (all fixations observed within a given area of interest; indicating how much time the participant spent reading the target area). To analyse eye tracking data, for each item, Areas of Interest (AoI) were defined by the researcher and a language teacher. These AoIs included text relevant key information in the text to answer a question and associated paragraph and question stem (QS). Text relevant key information included key words, phrases and sentences and paragraphs that had to be read to respond to a question and the relevant sentences were included in the analysis (See Appendix H, Table H1 and Table H2). This information was defined by the researcher and confirmed by a language teacher. One of the measures to be calculated on these areas is Total Fixation Count (TFC) which is defined as the frequency count of all individual fixations within a given AoI, that is how many times the target area was visited. The second one is Total Reading Time (TRT) which includes all fixations observed within a given AoI – indicating how much time the participant spent reading the target area. Rayner (1998) argues that it is better to measure TRT as a major EM measure when the target AoI is larger than a single word. In order to decide whether there were any differences in eye movements of participants across

parallel items depending on the formats open-ended and multiple-choice items, firstly, the reading time associated with the key areas was divided by the total reading time for the associated reading text. The rationale behind this was to measure test takers' attention and account for individual differences in their reading speed (Lim, 2014). After that, the test takers' eye movements on each MC item type (which are parallel in Text 1 and Text 2 MC) were compared to their counterparts in Text 1 and Text 2 OE item types in general and the eye movements of correct respondents in each format were also compared to see whether there were any differences based on item format. The eye movement data were then submitted to Mann-Whitney U test to see whether there were any differences in terms of test format effect on an item basis.

To answer the fourth research question, three types of strategy use will be investigated in the verbal report data which are reading strategies, test-management strategies and test-wiseness strategies. In order to analyse audio-recorded verbal report data, audio-recordings were first transcribed for each participant and then transcriptions were coded by the researcher according to reading strategies coding rubric, test-management strategies coding rubric and test-wiseness strategies coding rubric which were created based on the literature about reading and test-taking strategy use. However, after the first coding, it was clear that some segments that emerged in the data didn't correspond with the strategies in the rubric as each segment is to be seen as the representative of a single and specific process (Green,1998) and therefore new codes that emerged from the data were added to coding rubric. This was seen necessary taking into consideration what is suggested by Green (1998) about the fact that the small number of broad coding categories can lead to general and weak inferences. Therefore, coding was done in the way that each

segment in the data is met by a specific code. The verbal reports regarding MC and OE items were coded separately and twice. The first and second coding were completed by the researcher using the first version of the coding rubric. As a result of the first and second coding, in open-ended questions, there were 1149 codes in total and 957 (83%) of them were consistent while in multiple-choice questions, there were 1194 codes and 968 (81%) of them were consistent. These percentages ensured intra-rater reliability. However, the codes were checked for a third time and in the final version, there were 1096 codes in OE and 1140 codes in MC format and the discrepancies were resolved by more detailed analysis of the data or asking for an expert opinion. After the analysis was completed, it was observed that there were some overlaps or subtle differences in the coding rubrics and overlapping categories were combined with each other and a revised version was created for the coding rubric. The final codes were recoded based on the revised versions of coding rubrics for a fourth time and coding was finalized (See Appendix I for reading strategies coding rubric and Appendix J and Appendix K for test-management and test-wiseness coding rubric with their references respectively).

As a result of the codes, the frequencies for the usage of each strategy were calculated and compared in terms of overall reading strategy use and overall test-management and test-wiseness strategy use in both formats. In order to see whether there are any differences in terms of reading and test-taking strategy use in MC and OE formats, paired samples t-tests were conducted to compare the means of overall reading and test -taking strategy use (in terms of count) in OE and MC formats. Next, the most frequently used strategies were reported for each item type in the MC format only for correct respondents and they were compared with the most frequently used strategies in the parallel OE correct responses. The reason for this

was to explore the types of strategies used for each format on an item basis for correct respondents and understand what correct respondents do in each format to find the correct answer. It should be noted that this further analysis will be a more descriptive one.

Lastly, in the fifth research question, in order to understand the test takers' own perceptions regarding the item types, the interview data were qualitatively analysed. The interview data was analysed based on three themes that emerged from the data which are (a) test takers' perceptions about which format is more difficult (based on the tests they have taken), (b) test takers' own opinions about which item format was more helpful in their comprehension of the texts, (c) whether they have a specific preference for MC or OE questions and why. Their own reasons and justifications for their choice were discussed and direct quotations which were translated in English were included in the discussion part, too.

3.7 Conclusion

In this chapter, the research questions, participants profile, instruments, design and procedure of the study and lastly data analysis methods were explained in detail. The next chapter will explain the results that were attained through the data analysis done for each research question.

CHAPTER 4

RESULTS

4.1 Introduction

This chapter will explain the results of the data analysis. First, the results of the test performance comparison between the two formats will be reported to answer the first research question. Next, the analysis of eye movement data will be reported for each format to respond to the second and third research question. Then, the results of the strategy coding will be reported first as overall strategy use and secondly on an item basis for both formats for the fourth research question. Lastly, the results of qualitative analysis of interview data will be presented for the fifth research question.

4.2 The effect of test method on test performance

In order to answer the first research question which attempts to find out whether test format has any effect on test performance, it is first necessary to find out whether the difference in the scores resulted from method effect or text effect. Even though reading texts were chosen and adapted as parallel forms to each other based on a number of predetermined criteria, to make sure that the scores are not affected by the reading texts, first the descriptive statistics were calculated for both methods (Method 1-MC and Method 2-OE) and for two texts (Text 1 and Text 2) and they are presented in Table 7 and Table 8.

In terms of method, the results show that the mean (75.41%) of OE seems to be higher than the mean (60.82%) of MC and in open-ended questions, the scores are obviously higher in the second method, namely, open-ended format. However, for the test scores by text, it can be seen that the means (MC and OE combined) in each text are quite close to each other as it is clear in Table 8 and there is less variance.

To further assure whether the variance in the scores are affected by test method or text effect, a two-way analysis of variance was conducted taking score percent as the dependent variable and the method (MC or OE) and the text (Text 1 and Text 2) as categorical independent variables. The values of kurtosis and skewness (See Table 7 and Table 8) were within the acceptable levels (i.e., -2/ +2), suggesting that the scores based on the method and the text seemed to be normally distributed.

Table 7. Descriptive Statistics for Method 1 and Method 2

	Method 1	Method 2
Mean	60.82	75.41
Median	67	83
Std. Deviation	21.247	18
Skewness	-0.597	-0.659
Kurtosis	1.149	0.006
Variance	451.422	338.795

Table 8. Descriptive Statistics for Text 1 and Text 2

	Text 1	Text 2
Mean	65.68	70.56
Median	67	67
Std. Deviation	25.248	15.816
Skewness	-0.51	-0.311
Kurtosis	0.024	-0.356
Variance	637.438	250.133

The results of ANOVA are shown in Table 9 and they prove that the difference between the mean scores of the two tasks was due to test method effect. The result was not affected by texts used in the tests or an interaction of text and method. In addition, the effect for method was significant ($F= 5.394$, $p < .005$).

Table 9. Two-way ANOVA

Tests of Between-Subjects Effects

Dependent Variable: scorepercent

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	5993.059[a]	3	1997.69	5.394	0.002
Intercept	315520.94	1	315521	851.968	0
Method	3617.88	1	3617.88	9.769	0.003
Text	405.24	1	405.24	1.094	0.299
Method*Text	1969.94	1	1969.94	5.319	0.024
Error	23702	64	370.34		
Total	345216	68			
Corrected Total	29695.06	67			

a. R Squared = .202 (Adjusted R Squared = .164)

4.3 The comparison of eye movement data in both formats

In order to answer the second research question about whether there is a difference in the amount of texts that is carefully read depending on MC or OE format, Mann-Whitney U test was conducted to compare the percentages of careful reading times. In addition to that, text-based total reading time (TRT) and total fixation count (TFC) were compared to see if there are any differences regarding these measures, too. The results show that there is a statistically significant difference between open-ended and multiple-choice formats in terms of careful reading time, total reading time and total fixation count as it is summarized in Table 10. The results show that in open-ended format, the whole text is read for a longer time, there are more fixations in the overall text and a greater portion of the text is dedicated to careful reading.

Table 10. Eye-Tracking Statistics for Text-Based Careful Reading, TRT and TFC

		Text-based Careful Reading	Text-based TRT	Text-based TFC
MC (N=31)	Mean	75.76	200.46	503.3
	St. Dev.	9.58	64.16	114.9
OE (N=31)	Mean	80.36	243.57	613.06
	St. Dev.	12.25	78.54	158.46
Mann Whitney U		302.5	320.00	282.50
Z		-2.51	-2.26	-2.79
Sig (2-tailed)		.012	.024	.005

All significant at $p < .05$, n1 (MC) = 31, n2 (OE) = 31.

NOTE: Careful reading is in percentage, TRT in seconds, and TFC is the number of times.

To answer the third research question and to understand whether there are any differences in the eye movements, namely in TFC (Total Fixation Count) and TRT (Total Reading Time) of test takers in the areas of interest (AoI) - the parts of the text relevant to a specific question and question stem (QS) - in parallel items in both formats, the significance of differences between these two formats is calculated using Mann-Whitney U Test.

As it can be seen in Table 11, for the first item, text relevant areas (AoI) are read for a longer time and there are more fixations on these relevant areas in OE item. The difference between MC and OE in these measures are statistically significant. When it comes to total reading time and total fixation count in the question stem (QS), these areas are read for a longer time and fixated more frequently in OE item; however, there is no statistical difference between MC and OE in terms of QS TFC or TRT.

In Item 2, the same pattern as the first question is observed. Text relevant areas (AoI) in Item 2 are read for a longer time in OE item and fixated more frequently as AoI TFC and AoI TRT are higher in OE.

Table 11. Eye-Tracking Statistics for Item 1

		QS TFC	QS TRT	AoI TFC	AoI TRT
MC (N=31)	Mean	2.15	2.17	3.92	3.50
	St. Dev.	1.21	1.39	2.15	1.98
OE (N=31)	Mean	2.49	2.19	9.93	10.54
	St. Dev.	1.20	1.10	3.98	3.73
Mann Whitney U		379.00	425.00	84.00	44.00
Z		-1.43	-0.78	-5.58	-6.15
Sig (2-tailed)		0.15	0.44	0.00	0.00

This difference is statistically different as it can be seen in Table 12. As to the QS, total reading time in QS and total fixation count are higher in OE than MC, but there is no statistical difference between them.

Table 12. Eye-Tracking Statistics for Item 2

		QS TFC	QS TRT	AoI TFC	AoI TRT
MC (N=31)	Mean	2.06	1.95	6.52	6.64
	St. Dev.	0.94	1.13	2.45	2.61
OE (N=31)	Mean	2.67	2.63	9.23	9.63
	St. Dev.	2.10	2.23	3.34	4.29
Mann Whitney U		441.00	428.00	255.00	271.00
Z		-0.56	-0.74	-3.18	-2.95
Sig (2-tailed)		0.58	0.46	0.00	0.00

In Item 3, text relevant areas are visited more frequently and read for a longer time in OE item. In addition, total fixation count in QS is higher in OE. There is a statistical difference between MC and OE tests in terms of AoI TFC, AoI TRT and QS TFC. Although QS total reading time (TRT) is higher in OE, the result is not statistically significant. (See Table 13)

Table 13. Eye-Tracking Statistics for Item 3

		QS TFC	QS TRT	AoI TFC	AoI TRT
MC (N=31)	Mean	1.65	1.76	5.33	5.43
	St. Dev.	0.72	0.96	2.91	2.83
OE (N=31)	Mean	2.67	2.31	7.12	8.24
	St. Dev.	1.28	1.25	3.71	4.14
Mann Whitney U		234.00	355.00	328.50	263.00
Z		-3.47	-1.77	-2.14	-3.06
Sig (2-tailed)		0.00	0.08	0.03	0.00

In Item 4, total reading time in question stem is higher in MC than OE and the difference between them is statistically different. Total fixation count for question stem is higher in MC. As to the text relevant areas, the results show that total reading time and total fixation count are higher in OE than MC. However, the results are not statistically significant. For this item, the results show that in MC question, the test takers focused more on QS while in OE, they focused on text related areas. (See Table 14)

Table 14. Eye-Tracking Statistics for Item 4

		QS TFC	QS TRT	AoI TFC	AoI TRT
MC (N=31)	Mean	2.60	2.96	7.13	7.45
	St. Dev.	1.03	1.02	3.03	3.19
OE (N=31)	Mean	2.09	2.06	10.01	11.68
	St. Dev.	1.32	1.24	6.92	8.05
Mann Whitney U		343.00	288.00	402.00	359.00
Z		-1.94	-2.71	-1.11	-1.71
Sig (2-tailed)		0.05	0.01	0.27	0.09

As to Item 5, total reading time and total fixation count in text related areas are higher in MC than OE while test takers spent a longer time reading the question stem and visited the question stem more frequently in this OE item. For the QS TRT and

QS TRT, the differences between OE and MC are statistically significant. This result shows that students focused on QS more for this item in OE format (See Table 15).

Table 15. Eye-Tracking Statistics for Item 5

		QS TFC	QS TRT	AoI TFC	AoI TRT
MC(N=31)	Mean	2.03	2.08	8.21	8.63
	St. Dev.	1.47	1.60	2.85	3.07
OE (N=31)	Mean	2.93	2.71	7.23	7.76
	St. Dev.	1.50	1.53	3.96	4.73
Mann Whitney U		298.00	339.00	362.00	371.00
Z		-2.57	-1.99	-1.67	-1.54
Sig (2-tailed)		0.01	0.05	0.10	0.12

For the last item type, Item 6, students read the text relevant areas for a longer time and they visit this relevant part of the text more frequently in MC questions. The difference is statistically significant for these two measures. For question stem, while the total reading time and total fixation count are higher in MC than OE, the result is not statistically significant as it can be seen in Table 16.

Table 16. Eye-Tracking Statistics for Item 6

		QS TFC	QS TRT	AoI TFC	AoI TRT
MC (N=31)	Mean	1.69	1.81	7.91	7.81
	St. Dev.	0.85	1.19	4.18	4.08
OE (N=31)	Mean	1.63	1.54	4.64	4.59
	St. Dev	0.75	0.82	3.70	3.71
Mann Whitney U		463.00	440.00	243.00	240.00
Z		-0.25	-0.57	-3.34	-3.39
Sig (2-tailed)		0.81	0.57	0.00	0.00

The overall results show that in the areas of interest, total fixation count and total reading time tend to be higher in OE. This shows that the relevant parts of the text are read for a longer time and visited more frequently in OE tests (In 4 questions out

of 6). In a similar vein, in OE tests, question stem is read for a longer time and test takers visit this area more often (In 4 questions out of 6). The main findings of an item-based eye movement data are summarized in Table 17 regarding in which format total reading time is longer and total fixation count is higher in the areas of interest and question stem.

Table 17. The Summary Table for Overall Item-Based Differences

	QS TFC	QS TRT	AoI TFC	AoI TRT
ITEM 1	OE	OE	OE *	OE *
ITEM 2	OE	OE	OE *	OE *
ITEM 3	OE *	OE	OE *	OE *
ITEM 4	MC	MC *	OE	OE
ITEM 5	OE *	OE *	MC	MC
ITEM 6	MC	MC	MC *	MC *

NOTE: "*" stands for statistical differences at $p < .05$

A further analysis was conducted to compare only the correct respondents' eye movements in MC and OE format on item basis. The results for this comparison can be found in Table 18. As it can be seen in the table below, only for item 1 TRT and TFC increased for MC items. (For all the result tables of correct respondents for each item, see Appendix L, Table L1, L2, L3, L4, L5, L6.)

4.4 The comparison of reading and test-taking strategy use depending on test format

In order to answer the research question about what reading and test-taking strategies are used in different formats, verbal report data were analysed using the coding rubric and the codes are expressed in terms of their frequency for each format.

The results will first be presented as overall reading strategies and test-management strategies. Then, the results will be given on an item basis by comparing the strategies of both correct and incorrect respondents in two different formats.

Table 18. The Summary Table for Overall Differences for Correct Items

	QS TFC	QS TRT	AoI TFC	AoI TRT
ITEM 1	MC =OE	MC	OE *	OE *
ITEM 2	OE	OE	OE *	OE *
ITEM 3	OE *	OE *	OE *	OE *
ITEM 4	MC	MC	OE	OE
ITEM 5	OE *	OE	MC	MC
ITEM 6	MC	MC	MC *	MC *

NOTE: "*" stands for statistical differences at p< .05

4.4.1 Overall reading and test-taking strategy use

In order to respond to the research question about which reading and test-taking strategies are used in different formats, the frequencies were calculated and presented in Table 19. First, it can be seen that more reading strategies are used in open-ended items. The results of t-tests also prove that there is a significant difference in the reading strategy use in OE (M= 80.09, SD= 86.10) and reading strategy use in MC (M= 65.27, SD= 77.18) conditions; t(10)= -3.306, p= 0.008; d= 0.18. The effect size (d= 0.18) for this analysis was found to be low. This finding indicates that reading strategies are slightly more deployed in OE format. (See Appendix M, Table M1 and M2)

A closer look at Table 19 shows that prior to test taking, only 3.06% of participants read the text carefully before attempting the task in MC format while 2.72% of participants did the same in OE format (R1). None of the participants read the text expeditiously to have a general idea about the text (R2). The percentage of test takers who utilized expeditious reading strategies is slightly higher in MC format than OE format. Search reading (R6 + R4) is the top strategy for MC format which is followed by scanning (R3). Similarly, for OE items, search reading (R6 +R4) is the top strategy and it is followed by scanning (R3). The frequency of

skimming (R5) is very close in both formats. The strategy of reading only the parts of the text that seem relevant to the question and skipping the parts of the text seeming irrelevant (R8) is high in percentage in both formats; however, it is slightly higher in MC. With regard to careful reading strategy use, it can be seen that the percentage of careful reading in total is higher in OE than MC. The strategy of reading carefully across sentences by making connections between sentences (R10) is higher in OE than MC. The rest of careful reading strategies such as focusing on parts of a sentence to understand (R9), creating a textual representation (R11) and making inferences (R7) are slightly higher in OE, except for the strategy of rereading important parts (R14) which is higher in MC.

Table 19. The Frequency of Overall Reading Strategies

	MC	MC	OE	OE
	Count	Frequency	Count	Frequency
Prior to test taking				
R1 (Reads the whole text carefully before the test)	22	3.06%	24	2.72%
R2 (Reads the whole text quickly before the test)	0		0	
Expeditious Reading				
R3 (Scanning)	119	16.57%	120	13.62%
R4 (Search Reading)	3	0.42%	33	3.75%
R5 (Skimming)	33	4.60%	37	4.20%
R6 (Search reading)	144	20.06%	166	18.84%
	299	41.64%	356	40.41%
Other reading strategies				
R8 (Reading only parts that seem relevant to question)	254	35.38%	296	33.60%
Careful reading				
R9 (Focusing on parts of a sentence)	29	4.04%	42	4.77%
R10 (Reading carefully across sentences)	54	7.52%	87	9.88%
R11 (Creating a textual representation)	2	0.28%	3	0.34%
R14 (Rereading important parts)	29	4.04%	30	3.41%
R7 (Making inferences based on the text)	29	4.04%	43	4.88%
	143	19.92%	205	23.27%

As to the test-management and test-wiseness strategies, the frequencies of strategy use are summarized in Table 20 and Table 21 respectively. In order to see whether

there is a difference in the use of test-taking strategies in MC and OE format, a paired samples t-test was conducted and the results indicated that there is a significant difference in the test -taking strategy use in OE (M=11.38, SD= 17.98) and test-taking strategy use in MC (M= 21.10, SD= 20.96) conditions; t (20) = 2.159, p= 0.04; d= 0.49. The effect size (d= 0.49) for this analysis was found to be moderate. (See Appendix N, Table N1 and N2 for t-test results).

In terms of the test-management strategies, for OE items, the most frequently used strategy is going back to question for clarification with the help of rereading, translating, etc (T1) and the second most used strategy is using the order of questions to locate the text span where the answer is located (T6). The next most common strategy is stopping reading the text when they think the answer is found (T36) or producing their own answer after reading the relevant parts of the text (T8). The strategies which are close to each other in terms of frequency are continuing to read the text when the answer is found (T35) and skipping the difficult question mostly to save time (T5). The least common strategies are expressing uncertainty (T29) and receiving clues from other items to answer a question (T31). Identifying an option with unknown meaning (T11) or selecting preliminary options with uncertainty (T14) are used in OE vocabulary questions, but they have low frequencies. Similarly, making an educated guess (T19) is one of the lowest strategy.

As to the multiple- choice questions, there are a greater number of test-management strategies used in MC than in OE. There appears to be a separate category among test-management strategies which is “option-related strategies” which is presented in Table 22 and 43.49% of test-management strategies in MC items is option-related while in OE items, only 2.98% of strategies are option-related and they are used in vocabulary questions. In MC, test takers go back to questions or

options to clarify the meaning (T1) and this strategy is lower than in OE test regarding both its ranking and percentage. Producing their answer after reading the text (T8) is the next one and this is the same as OE in terms of ranking and similar to OE in percentage. The next finding is that the test takers stopped reading options (T30) when they reached an answer. T14 and T19 are used more frequently in MC than OE. Next, test takers read the questions and the options before reading the text (T4).

Table 20. The Frequencies of Test-Management Strategies

	MC Count	MC Frequency	OE Count	OE Frequency
T1 (Rereading question for clarification)	42	9.77%	66	28.09%
T4 (Reading question and options before the text)	32	7.44%	0	
T5 (Skipping a difficult question)	13	3.02%	18	7.66%
T6 (Using the order of questions as a clue)	64	14.88%	53	22.55%
T8 (Producing answer after reading the text)	42	9.77%	22	9.36%
T11 (Identifying an option with unknown vocabulary)	5	1.16%	3	1.28%
T14 (Selecting preliminary options with uncertainty)	34	7.91%	4	1.70%
T17 (Eliminating similar options)	2	0.47%	0	
T18 (Wrestling with option meaning)	3	0.70%	0	
T19 (Making an educated guess)	33	7.67%	4	1.70%
T22 (Selecting options based on background knowledge)	6	1.40%	0	
T23 (Selecting options based on paragraph meaning)	70	16.28%	0	
T29 (Expressing uncertainty at the correctness of an answer)	23	5.35%	14	5.96%
T30 (Stopping reading the options when the answer is found)	35	8.14%	0	
T31 (Receiving clues from other items)	6	1.40%	9	3.83%
T35 (Continuing to read the text when the answer is found)	6	1.40%	19	8.09%
T36 (Stopping reading the text when the answer is found)	14	3.26%	23	9.79%

Table 21. The Frequencies of Test-Wiseness Strategies

	MC Count	OE Count
TW1 (Selecting option out of a vague sense, even if it is not understood)	3	0
TW2 (Using clues in other items to answer the item under consideration)	2	2
TW3 (Selecting an option as it has a key word / phrase from the passage)	8	0
TW4 (Chooses a phrase as answer which is in the same sentence as the key word)	0	2
	3%	1%

Table 22. Option-Related Strategies

	MC Count	MC Frequency	OE Count	OE Frequency
T11 (Identifying an option with unknown vocabulary)	5	1.16%	3	1.28%
T14 (Selecting preliminary options with uncertainty)	34	7.91%	4	1.70%
T17 (Eliminating similar options)	2	0.47%	0	
T18 (Wrestling with option meaning)	3	0.70%	0	
T23 (Selecting options based on paragraph meaning)	70	16.28%	0	
T22 (Selecting options based on background knowledge)	6	1.40%	0	
T30 (Stopping reading the options when the answer is found)	35	8.14%	0	
T4 (Reading question and options before the text)	32	7.44%	0	

The low-frequency strategies can be seen as expressing uncertainty at the correctness of the answer (T29), skipping a question when it is difficult (T5), and stopping reading when the answer is found (T36). Eliminating similar options (T17) and wrestling with option meaning are the least common strategies in MC format (T18).

With regard to the test-wiseness strategies, it is obvious in Table 21 that test takers attempt to use test-wiseness strategies more in multiple-choice questions than open-ended items and the top strategy in MC is choosing an option merely because it has a key word from the text without reading the relevant parts of the text (TW3). This strategy is followed by elimination of items and choosing an option out of a vague sense that other options cannot be true (TW1) and the last one is using clues in other items to answer an item (TW2). While the percentage of test-wiseness strategies is higher in MC, there are also attempts of test-wiseness strategies in OE items. There is a test-wiseness strategy that is specific to OE items and it is when test takers choose a phrase or word as answer only because it is in the same sentence as the key word without considering about its correctness (TW4).

4.4.2 Comparison of reading and test-taking strategy use on an item basis

In this part, the item types which are parallel to each other in Text 1 and Text 2 (See section 3.4.1, Table 3.4) in MC format will be compared to their counterparts in OE format to describe the types of strategies used in each item regarding the most frequently used strategies in correct respondents. The results will be presented in tables for each item and will be followed with a short explanation.

For the first item which is a local reading question, as it can be seen in Table 23, the top 3 strategies in correct respondents' MC and OE items are similar as both include reading relevant parts (R8), search reading (R6) and scanning (R3). However, OE also involves reading carefully across sentences (R10) while MC includes elimination of options based on the options (T23) as a fourth strategy.

Table 23. Item 1 Factual/ One sentence

	MC			OE		
	Strategy	Count	Frequency	Strategy	Count	Frequency
Correct	R8	17	29%	R8	43	27%
	R6	12	20%	R6	32	20%
	R3	7	12%	R3	14	9%
	T23	5	8%	R10	13	8%

For Item 2, for correct respondents, the top 3 strategies are similar again making use of reading relevant parts (R8), search reading (R6) and scanning (R3). For the correct respondents of OE items, in addition to these, there is the strategy of reading carefully across sentences by making connections (R10) while in MC, the fourth most frequent strategy is using the order of questions as a clue to locate the text span where the answer is located (T6). (See Table 24)

Table 24. Item 2 Factual/ One paragraph

	MC			OE		
	Strategy	Count	Frequency	Strategy	Count	Strategy
Correct	R8	34	22%	R8	41	26%
	R6	23	15%	R6	30	19%
	R3	17	11%	R3	19	12%
	T6	15	10%	R10	14	9%

In Item 3, which is a vocabulary question, correct respondents make use of different types of strategies in both formats. In MC format, after reading relevant parts (R8) which is the top strategy, respondents make an educated guess based on their background knowledge (T19), they focus on parts of a sentence (R9) and the least frequent strategy is reading carefully across the sentences (R10). However, for OE format, to answer vocabulary questions, respondents read relevant parts of the text and mostly focus on parts of the sentence (R8 and R9 respectively). They read carefully across sentences (R10) and use search reading strategies (R6) (See Table 25 for the frequencies).

Table 25. Item 3 Vocabulary

	MC			OE		
	Strategy	Count	Frequency	Strategy	Count	Frequency
Correct	R8	21	24%	R8	39	35%
	T19	18	20%	R9	14	13%
	R9	12	14%	R10	11	10%
	R10	8	9%	R6	9	8%

In Item 4, it is clear that correct respondents in MC format mostly read only relevant parts of text (R8), they use search reading and scanning (R6 and R3 respectively). They make use of the order of questions to define relevant parts of the text to read

(T6) and lastly, they select the options based on an elimination of other options. In OE format, correct respondents make use of the same strategies as MC format except for the last strategy which is reading carefully across sentences (R10) (See Table 26). As to the item 5 which is an inference question, it is observed that in MC format, correct respondents utilize the strategies of reading relevant parts (R8) and scanning (R3). (See Table 27)

Table 26. Item 4 Factual/Across sentences

	MC			OE		
	Strategy	Count	Frequency	Strategy	Count	Frequency
Correct	R8	37	28%	R8	46	27%
	R6	19	14%	R6	26	15%
	R3	10	7%	R3	25	15%
	T6	10	7%	T6	16	9%
	T23	7	5%	R10	11	7%

Table 27. Item 5 Inference / Paragraph level

	MC			OE		
	Strategy	Count	Frequency	Strategy	Count	Frequency
Correct	R8	24	24%	R8	36	33%
	R3	9	9%	R10	14	13%
	R6	6	6%	R7	13	12%
	R10	6	6%	R6	8	7%
	R7	6	6%	T1	8	7%

The strategies that are used in item 5 in the same percentage are search reading (R6), reading carefully across sentences (R10) and making inferences based on the text (R7). In open-ended format, correct respondents read relevant parts (R8), read carefully to make connections (R10) and make inferences based on the text (R7). The least frequent strategies for this question is search reading and going back to question for correct respondents in OE format.

In Item 6, as it can be seen in Table 28 clearly, correct respondents in MC questions mostly read only parts of the text that seem relevant to the questions (R8), search reading (R6) and scanning (R3) and these are followed by producing their own answer after reading (T8) and choosing the correct option by eliminating other options based on the text (T23). When it comes to the OE question, it is seen that the top strategies are reading relevant areas (R8), scanning (R3) and search reading (R6) respectively and these are followed by reading carefully by making connections (R10) and going back to the question for clarification (T1).

Table 28. Item 6 Factual/ Across paragraphs

	MC			OE		
	Strategy	Count	Frequency	Strategy	Count	Frequency
Correct	R8	29	20%	R8	36	23%
	R6	28	19%	R3	27	17%
	R3	18	12%	R6	23	15%
	T8	10	7%	R10	15	10%
	T23	9	6%	T1	11	7%

4.5 Test takers' own perceptions about test method effect

In order to answer the research question that attempts to explore test takers' own perceptions about test method effect, the results of the interviews are categorized under three themes that emerged in the data. These categories are (a) which format is more difficult; (b) which format they comprehended better; (c) which item format they prefer to respond. The results are calculated based on the frequencies and presented in Table 29.

Fourteen participants stated that MC questions were more difficult for them and when the reasons for this were asked, they mostly mentioned that the

distractors confused them and in MC questions, they felt themselves limited in terms of producing answers.

Table 29. Results of Interviews

Difficulty	Comprehension	Overall Preference
14 MC- 43.75%	28 OE – 87.5%	17 OE – 53.12%
14 OE – 43.75%	3 BOTH – 9.37%	12 MC – 37.5%
2 BOTH- 6.25%	2 MC – 6.25%	3 NA – 12.5%
3 NONE- 9.37%	1 NA -3.12%	2 BOTH – 3.12%
1 NA - 3.12%		

NOTE: NA stands for "No Answer"

They also stated that they were more passive while answering MC questions and they could answer some questions without fully understanding the text. Fourteen participants stated that OE items are more difficult for them as they don't have any options through which they can have an idea about the question intent and they had to understand both the text and the options. Three participants reported that both were equal in terms of difficulty. Twenty-eight participants thought that they comprehended the OE text better and among the reasons for these, they stated that the fact that they had produced their own answers with their own words. This made them read the text in a more careful and detailed way. In OE, they reported that they had to reread some parts to produce an answer. Some claimed that even though OE test was more difficult, they understood it better and answered with more confidence. Three participants stated that they understood the texts to an equal extent. Interestingly, two participants whose scores were higher in OE reported that they understood MC text more. Seventeen participants reported that they preferred answering OE items as that format doesn't confuse them during their reading process and they enjoy it more regarding overall preference. Twelve participants stated that although they understand OE more, they prefer MC items more as it is more practical to answer MC items and they can have clues from the options. Two participants

remarked that test format doesn't matter as what matters is the content of the questions. One of them mentioned that in either case, 'You need to comprehend the text first before answering the questions'.

Translations of direct quotations from the interviews together with participant number will be presented below.

Participant 39

"Well, in MC questions, I answered questions by eliminating the options and by using key words. In the second one, there was only key word but no elimination (of options). I had to produce my own answers and it was difficult for me. In MC, you can at least guess from the options. In MC, you can guess the answer more easily, the answer is among the options…"

Participant 30

"Even if I don't understand the paragraph, I reached an answer (talking about MC items). I could find an answer without understanding. But in the second one (OE), there are no options but you have to concentrate and understand (the text) to come up with an answer. It also depends on your ability to interpret…. The second one (OE) requires more effort."

Participant 27

"I think test format effect has an effect (on comprehension) because I was more relaxed in open-ended items. The questions were given and I produced my own answer by reading the relevant parts. However, in MC, I felt more… like there is only one answer and I have to find it in the options. In OE, I was more flexible. I thought I can read and form my own sentences. In MC, there was no flexibility and it limited me…"

Participant 22

"I feel like MC was more difficult for me because I feel it limits me. Before reading the options of the questions, I feel afraid… what if the answer I have given is not among the options? What if I have understood everything in a wrong way? I feel nervous and limited. But in OE I have the freedom of reading and making sense out of the text on my own and interpret it as I wish. Therefore, I feel more comfortable while answering OE questions…"

4.6 Conclusion

This chapter presented the findings regarding each research question in a detailed and comprehensive way. First, the statistical results were presented to show the test method effect on test performance. Second, eye movement data were analysed to show differences in how much of the text is read carefully in each format and how total reading time and fixation counts with regard to interest areas (namely question stems and parts of the text relevant to each question) differ for each item in both formats. Thirdly, the frequencies of reading and test-taking strategies were given first in an overall basis comparing methods with each other and then to have a further insight into the processes, read they were compared on an item basis across methods. Lastly, the interview results were presented in terms of frequency of three themes along with translations of direct quotations from the participants.

CHAPTER 5

DISCUSSION

5.1 Introduction

Investigating the cognitive processes that a test taker undergoes while taking a specific test can be useful as it can reveal the processes that they have undergone and show the processes through which they come up with a specific answer. Test writers are required to show that the items they have written can elicit the intended processes from test takers. Otherwise, the inferences and the decisions that have been made on these scores cannot be claimed to be valid. As in some cases, it might be possible that test taker gets a full score from a test without understanding the text in a real sense or going through the processes intended by the item writer. In this context, this test cannot be claimed to be a cognitively valid one. As the processes can be more informative than the scores most of the time, a process-based validation is preferred rather than the product (score)-based validation. When it comes to the reading ability which is mostly tested by indirect test methods, the necessity of understanding these processes becomes more crucial. When the test format effect is taken into consideration, the issue of strategy use becomes a subject to pay attention to. Test takers can utilize some strategies, which can be both relevant and irrelevant to the construct, to cope with the item demands. These strategies that are irrelevant to the construct undermine the validity of a test. To this end, the current study attempted to compare two test methods, namely multiple-choice and open-ended items, in terms of the reading processes and strategies they trigger through the use of retrospective verbal reports, eye-tracking methodology and short semi-structured interviews.

The rest of the chapter will discuss the findings from different sets of data in relation to the research questions. Each subsection will focus on a different set of data that attempt to respond the relevant research questions. There will be a section in which the results of correct respondents' eye-tracking data and verbal reports will be discussed on an item basis to see whether test takers who get correct scores go through the same processes while responding to the same item type in different formats.

5.2 The effect of test method on reading performance

To answer the first research question which attempts to find out whether differences in the test format, namely MC or OE formats, can affect the scores in the reading test, the two reading texts chosen for the current study were comparable in terms of a number of criteria and the items were created parallel to each other regarding item type and text span (See Methodology section). The underlying reason for this is that the study aims to understand the pure effect of test method and whether it would affect the scores. The results of the tests indicate that the scores in OE are obviously higher than MC. To see whether the difference in the scores stems from the method or text, a test of ANOVA was conducted to see whether there is a method effect on the scores and the results reveal that test method affects the scores and the variance in the scores can be explained by the method effect most. The results regarding method effect are seen to be statistically significant ($F= 5.394$, $p < .005$). This finding deviates from the findings of previous studies (Lim, 2014; In'nami and Kouzumi, 2009) which show that test takers performed better in MC than OE. In'nami and Kouzumi assert that some conditions should be met to assure that MC is higher than OE and these conditions include between-subjects design, random assignment, stem-equivalent items and advanced learners. Although two of these

conditions – stem-equivalent items (also content-equivalent items) and advanced learners- are met in the study, the scores in OE are higher than MC. The reasons underlying this difference in the scores can be explained by the differences in the cognitive processes, reading and test-taking strategy use which are elicited by different test methods which will be discussed in the subsequent research questions in detail.

5.3 The results of eye-tracking data

The results of eye-movement data are used to answer the second and third research questions. To respond to the second research question, the careful reading percentages, text-based total reading time and text-based total fixation count were compared in MC and OE format using Mann-Whitney U test and the results of the test prove that all of the measures, namely the percentage of careful reading, text-based total reading time and text-based total fixation count, are used in the texts with OE items more and the difference between the texts in OE and MC formats in terms of these three measures is statistically significant. This suggests that in open-ended items, a greater percentage of the text is carefully read when compared to multiple-choice items. This is an expected result when the higher scores that the participants got from the open-ended tests are considered. They spend a longer time to process and understand the text, they fixate on the text more often as it is expected from the successful readers. All the items in the tests used in the current study target at eliciting careful reading abilities and in this context, it can be claimed that open-ended items are more successful in making the test takers become engaged in the intended reading processes. This result indicates that test takers spent a longer time in the text to understand, to process and to make sense out of the text in OE. As the

scores in MC are lower, the shorter TRT doesn't seem to signal processing efficiency. The reason why the percentage of careful reading is lower in MC might be explained with the fact that either test takers focused more on options to answer the questions and they were distracted by the options or they used expeditious reading skills such as search reading, scanning or skimming to respond to the questions which shortened the careful reading time. It is probable that instead of reading the sentences carefully in a linear way, they might have matched the key words that appeared in the options with the text or searched the text to find the words in the options without substantial comprehension. As a result, it can be asserted that they went through different reading processes in MC test when compared to OE test. The results in this study for this question are in line with Lim (2014) who suggests that test takers spent more time reading a text, a question stem and key phrases when the OE test is given and OE test as a result may demand deeper cognitive processing which requires more complex cognitive processes. The fact that test takers had to produce their own answer might have made them focus on the text in a more careful way and the absence of options in OE might have encouraged them to read the text more linearly.

The third research question investigated whether there were any differences in total reading time (TRT) and total fixation count (TFC) on an item basis between two formats. For OE items 1, 2, 3, 4, the test takers focused on text relevant areas for a longer time and visited these areas more often while in questions 5 and 6, the measures of total reading time and total fixation count in text relevant areas are higher in MC. As to the question stem, the measures of total fixation count and total reading time are higher in OE in questions of 1, 2, 3, 5 while the same measures regarding the question stem are higher in MC in the questions of 4 and 5. Item-based

results confirm the overall picture regarding the results careful reading percentages and demonstrates that OE test tends to lead the test takers to spend more time in the related areas of the text and make them visit these areas more often. This finding can be interpreted as OE items make respondents focus on reading and understanding the text more than MC items. In addition, it is also possible to argue that in OE items, test takers can locate the information related to a question more easily, most probably as a result of careful and linear reading. They could realize the importance of the relevant areas easily and spend more time in the relevant areas to process and understand them. However, in MC items, they might have located the information wrongly as a result of scanning to match the key words from the text with the options, which is suggested by shorter reading time in the relevant areas of the text. It might also be argued that the exact key word matching between options and the text might have lead the students to look for the response of the answer in an irrelevant area. Due to the lack of careful reading, they might have focused on every area where there is a key word and they might mistakenly have read the wrong areas and tried to respond based on them. Therefore, eye-movement data for the items suggest that in OE test takers can locate the relevant information more easily as a result of careful reading and they might have focused on these areas more to find the answer. Their reading time and fixation counts in question stem in OE test shows that they read to understand the questions.

A further comparison in the eye-movement data of only correct respondents in terms of their total reading time and total fixation count in the text relevant area and question stem shows that correct respondents in OE still read the relevant areas of the text for a longer time when it is compared to MC condition. The fact that correct MC respondents spent a shorter time in related areas suggests that either

correct respondents in MC answered the questions by key word matching between the text and options or they gave a correct answer by reading irrelevant areas of the text and using the clues in the options. They might have depended more on their expeditious reading skills. As to the question stem, the results are mostly close to each other in both formats in terms of total fixation time and count. However, in vocabulary and inferencing questions, total fixation count is higher in OE with statistically significant results. This suggests that correct respondents in both formats try to visit the question stem and focus on it to a relatively similar extent.

5.4 The results of verbal report data and semi-structured interviews

Verbal report data are collected from test takers to see what kind of reading and test taking strategies are used while answering questions in different formats. The results of strategy use will be discussed by referring to the results from the interviews.

First of all, in terms of the number of strategies used, it is possible to say as indicated by the overall number of strategy use, test takers mostly used reading strategies, followed by test-management strategies and the least used strategy is test-wiseness strategies. The reason for this can be explained with the fact that all the participants in the study were advanced learners of English and it is understandable that they don’t need to use as many test-taking strategies as reading strategies. This overall finding is also consistent with Cohen and Upton (2006) in that they also claim that regardless of the differences in task types, overall test takers draw on the strategy of comprehending the meaning. The test takers in this study also engaged in some test-management strategies to respond to test items and they reported fewer test-wiseness strategies. The results of the tests also reveal that in MC, the test takers reported more test management strategies while in OE, they reported more reading

strategies. This result also confirms eye-tracking results with regard to the fact that the text in OE format is read more carefully and more time is spent to read the text in OE format.

As to the reading strategies, it is observed that except for four test takers, all the test takers first read the question and then went to the relevant area in the text. It is seen that both groups made use of expeditious reading skills. The test takers' behaviours in OE and MC differed from each other in terms of types and frequencies of careful reading they utilized. They were more engaged in careful reading in both local and global level in OE format. This explains why there were higher scores in OE format and also confirms eye-tracking results. As they read more carefully in OE, it can be seen in the current study that even for the same item type, they give more attention to the key information in the text (Lim, 2014) and they engage in a more complex thinking in OE (Martinez,1999) by making connections while reading carefully across sentences.

When it comes to the test-management strategies, first it should be noted that in MC, almost 43% of total test management strategies are option-related. Based on my personal observations and the interview data, it can be asserted that options can help test takers find the correct answer to some extent. However, they might also mislead the test taker. Some of the test takers remarked in the interviews that even if they understood the question intent, they read and understood the relevant area, they sometimes gave a wrong answer because of the options. They said that it is sometimes difficult to match the meaning you get from the text with the meaning in the options. It was difficult for them to understand someone else's comprehension as stated in the options and this made them nervous. In addition, when they thought they found the correct answer and stopped reading the options, they could choose the

wrong answer. Most of the time, even if the correct option was among their preliminary options, the options could mislead them due to their tricky nature or subtle differences among them. This leads them to purposeful reading behaviours and they start to be more strategic to cope with the item demands. They try to engage in additional processes and problem-solving strategies. These attempts can even result in test-wiseness strategies in which test takers come up with an answer without reading or understanding the text properly. As previously mentioned, the attempts of test-wiseness strategies was very limited in the current study (MC= 3%) as the participant group consisted of advanced learners of English and they strived to do their best. What should be considered here is that when the attempts of test-wiseness strategies result in correct responses, the validity of the test is seriously undermined. As it is also pointed out by Field (2011), these strategies specific to the MC format can undermine the test's validity; all of these strategies are very format specific strategies and it was observed that they disrupted the linear reading process of test takers. Paulson and Henry (2002) suggest that completing cloze items alters normal reading processes. In a similar vein, completing MC items in this study by the attempts to match key information in the text with the options or question rather than continuous reading can be claimed to alter normal reading processes of test takers to a greater extent in comparison to OE format. The results in this study are also consistent with the results of Rupp, Ferne and Choi (2006), who also suggested MC and OE trigger different cognitive processes. Reading a text with MC questions is seen as a problem-solving activity to cope with item demands to find the correct answer. They further claimed that responding to MC questions is never a linear process and reasoning in MC is a unique process. These claims overlap with the findings of the current study and the probable reason for this segmented reading is

the use of option-related strategies. It is observed that as when the test takers had options, they tried to match the key words from the text with the options most of the time and reading time was shorter, which suggested that they made use of faster reading skills without reading the relevant parts in a slow, careful and continuous fashion.

It is also observed in the data that there are some test-management strategies in OE items, too. The top test-management strategy is going back to question for clarification by rereading the question or translating it. When the paragraph number is not given in the question, the test takers in this study tried to use the order of questions to locate the text span. They stopped reading the text when they thought they found the correct answer. Along with this strategy, the test takers expressed the certainty at the correctness of their answer. In other cases, they reported that they wanted to continue reading to make sure that they found the correct answer. Therefore, the test-management strategies used in OE items cannot be claimed to be as disruptive as the ones used in MC items in terms of changing the expected reading processes and leading test takers to wrong answers due to confusion.

The interviews also confirmed these results as some participants reported the difficulties they had in MC stemmed mostly from the options. They also claimed to have comprehended the text in OE format better and the reason was that as they had to find the answers themselves, they read the text in a more focused way and they tried to understand the sentences more. Some reported that as the next step is writing the answers with their own words after reading the relevant areas, they needed to understand what they read very well. Some of them also reported to have answered OE questions more easily. In terms of difficulty, the number of test takers reporting that MC was difficult was the same as the number of the ones who reported OE was

difficult. In MC, they reported that options can make the task difficult while in OE producing their own answers made the OE format more difficult for them as it required more comprehension and effort.

5.5 Item by item comparison between eye movement and verbal report data to investigate cognitive processes depending on format effect

In this part, the results of the eye-movement data as measured by total fixation count and total reading time will be discussed by making connections with the participants' reading and test taking strategy use as indicated by their verbal protocols for only correct responses. The reason why there is such a discussion part is to make a direct and systematic comparison between MC and OE correct respondents and to see how they differ in each item format in terms of their cognitive processes. The goal is not to show which strategy is used more or less at this point but to identify the types of strategies used and to see whether they make use of different or similar strategies in different formats. In addition, correct respondents were expected to spend more time in the relevant area in the text with more frequent visits but this was not the case for most MC items. Therefore, only such a systematic comparison will be helpful in understanding the differences in the processes. At the beginning of the study, the importance of understanding cognitive processes is highlighted and it is highly important to see whether the parallel items trigger similar processes or not in correct respondents. Discussing eye-movement data and verbal reports together is expected to yield more illuminating results.

The results of eye movement data show that in Item 1, the relevant area is read for a longer time and it is visited more in OE and it can be seen in the verbal reports that test takers in MC reported to have used more expeditious reading skills

such as search reading and scanning and additionally, they eliminated the options based on the text. This can show that correct respondents in MC depend more on matching the words to answer a question rather than reading carefully. However, in OE, test takers reported to have read the text carefully along with the expeditious reading, namely search reading and scanning. This may indicate that after locating the information quickly, they read the relevant area carefully to find the answer. In this item, the verbal reports confirm eye movement data as test takers also claim that they read the text more carefully in open-ended format.

For Item 2, eye-movement data show that in open-ended questions, the test takers fixated on the relevant area in the text more often and for a longer time. When the verbal reports are considered, in OE, the test takers reported to have made use of search reading and scanning in general along with careful reading across sentences. However, in MC questions, the test takers reported that they made use of search reading, scanning and using the order of questions to locate the text span to find the answer. Here, it should be noted that as the test takers in MC didn't read carefully, they might have answered the question more superficially by matching the key word. The order of the questions sometimes misled them and they found the correct answer based on an irrelevant paragraph by option elimination or matching words in the text with the options.

Item 3 is a vocabulary question and eye movement data suggested that the test takers read the relevant area in the text for a longer time and they visited this area more often in OE items. These results are consistent with verbal report data in that test takers reported that they read the parts of a sentence carefully and they read across the sentences carefully to understand the context. This shows that they spent more time in the relevant area in OE format. However, in MC items, the test takers

reported that they made an educated guess by depending on their background knowledge along with reading parts of a sentence and careful reading across sentences. However, their reading time is quite short as shown by eye-movement data. Therefore, it is possible that they didn't get the meaning of the word from the context in MC as opposed to the OE format. Therefore, the validity of MC vocabulary item can be seen as questionable.

Item 4 shows that correct respondents in OE items read the relevant areas for a longer time and visited these areas more in comparison to the MC items; there was not a significant difference in the reading times for this question. The verbal report data indicated that in MC, the test takers made use of search reading, scanning and eliminated the options and located the information by paragraph number. However, in OE item, the test takers reported that they also used search reading and scanning but along with careful reading. For this question, careful reading is also utilized by the test takers. However, some students in OE also reported that they searched quickly for key word to answer the question.

In Item 5, correct respondents in MC read the relevant areas of the text for a longer time and they visited those areas more often. This question requires the test takers to make some inferences based on the text. In MC, the test takers reported that they used more reading related strategies such as search reading, scanning, careful reading and making inferences. For this question, in OE format, the test takers also reported that they made inferences based on the text and read the relevant areas more carefully. It might be the case that the correct respondents read the relevant paragraph where the answer is located carefully before this question while responding to the former question 4, they read the paragraph for a second time. Therefore, it can be claimed that the fact that their reading time is lower doesn't

mean that in OE, they didn't read the paragraph carefully as they could have processed it more quickly based on what they remember from the text for question 4. They still utilized their inferencing abilities together with careful reading. However, as to the correct MC respondents, because they didn't read the paragraph carefully for the previous question (question 4) and they needed to read it to make inferences, their total reading time is higher as they couldn't match the key words in the options and the text as they needed inferencing.

In Item 6, eye movement data indicate that the relevant area in the text is visited more and total reading time is longer in MC questions. The verbal reports show that correct MC respondents claimed that they made use of scanning, search reading, producing their own answers by reading the relevant part and eliminating the options. This shows that the longer reading time and more frequent visits can be explained with the fact that the test takers tried to match either the key words in the questions / options with the words in the text. This process of going back and forth might have increased the total reading time. It should be also noted the question stems were longer including several key words and test takers had more words to search for. As to the OE items, based on my personal observations as a researcher while listening to the test takers' verbal reports, most reported a more linear process as they located the part where the answer is and read that part carefully and then they responded and gave their answers with some certainty. In addition, some test takers also claimed that overall text meaning helped them give an answer, too. Considering the fact that they read a greater portion of the text carefully as indicated by both verbal reports and eye movement data by the time they have arrived the last question, it is possible to understand why they could respond to this question in a shorter time than the ones MC.

When the levels of cognitive processing in reading tests are considered (Khalifa & Weir, 2009), even with such a limited number of items that are parallel in MC and OE formats, it is possible to see that MC items trigger lower levels of cognitive processing such as word matching or word class / synonym matching as shown by the dependence of MC items on scanning and search reading that are supported by test-taking strategies more. They resort to higher levels of cognitive processes only when they cannot answer by using test-management strategies. However, in open-ended formats, test takers utilize higher cognitive processes in addition to lower levels of cognitive processes such as reading carefully across sentences by making connections, making inferences based on the text or building a mental model of the text. Some students in Item 6 reported that they couldn't have answered that question if they had not read the whole text and they reported a textual level of understanding. The results suggest that OE items make students read the text more carefully. These results confirm the results by Rauch and Hartig (2010) that while MC items correlate with basic reading processes, OE items correlate with both basic and higher reading processes. This is also the case in this study which shows that there is a wider range of reading processes used in the OE test. In addition, the fact that MC vocabulary questions could be solved by background knowledge rather than reading and guessing from the context overlaps with the findings of Lim (2014). In that study, the test takers could answer the vocabulary question without using the context as a clue and this undermines the validity of MC vocabulary items when compared to their OE counterparts which made the test takers read, understand and answer according to the context.

Based on the item-by-item discussion in this subsection, as it was discussed at the beginning of the study, the importance of investigating cognitive processes that

test takers go through by identifying their reading types, their reading and test-taking strategies and understanding what they do when they are faced with specific tasks becomes clearly evident as these processes have so much to tell us about what happened. Specifically, even the comparison of processes of correct respondents revealed many things that couldn't be understood by merely looking at the scores. It is also worth noting the complementary and explanatory effect of eye-tracking technology when combined by retrospective verbal reports in investigating how test takers respond to a task.

5.6 Conclusion

This chapter has discussed the main findings of the current study with the help of different types of data by both attempting to discuss and interpret their meaning on their own and in relation to other studies. The next chapter will focus on the conclusions drawn from this study along with the limitations and implications of the study.

CHAPTER 6

CONCLUSIONS

The current study was motivated to find whether there were any differences between open-ended and multiple-choice format in parallel reading comprehension questions in terms of the cognitive processes, reading and test-taking strategies that test takers utilize. In order to investigate this difference, five research questions were asked. The results indicated that the test takers had higher scores in open-ended tasks. As shown by the eye movement data, text-based careful reading percentage, total reading time and total fixation count were higher in OE format when compared to MC format. It was seen that the question stem and related areas in the text are read for a longer time and fixated more often in OE in 4 items out of 6. Verbal report data indicated that while in MC, test-management strategies appear more, in OE reading strategies are used more frequently. A deeper analysis showed that option-related strategies are mostly preferred in MC items. Then, for each item, a comparison was made between MC and OE items to find out the cognitive processes the participants went through with the help of eye-tracking and verbal report data. Lastly, the majority of test takers reported to have comprehended OE texts and items more in the interviews.

In the light of the evidence presented in this study, it can be concluded that test method causes a difference in the cognitive processes, reading and test-taking strategy use as proven by verbal reports, eye-movement data and interview results. It can be concluded as a result of both quantitative and qualitative evidence in the study that open-ended tasks necessitate or facilitate more careful reading and test takers focus on relevant areas in the text more. Therefore, limitations of MC format should be considered when this format has to be used. MC items can confuse some test

takers and can have an effect on their scores. Additionally, in MC, linear and careful reading is disrupted more due to the options. Options can help the test takers, but it can also mislead them in some cases. Therefore, test writers should pay attention to the quality of each option. Options should not be based on absurdities, ambiguities or subtle differences but rather on the comprehension of the reading text. Another point is that MC items increase the number of test-management strategy use while OE items trigger the use of reading strategies more. Test-taking strategies are expected to be used in MC format more as the nature of this format triggers this. In OE, test-management strategies are also used but they support the reading comprehension process more and disrupt the reading process less than the ones used in MC. As to the reading strategies, expeditious reading is observed both in MC and OE as test takers try to locate the part where the answer is found in the text with the help of scanning or search reading. However, in OE format, expeditious reading is supported more by careful and extended reading when compared to the MC format. The fact that the text is carefully read in OE shows its effects on scores and this is also proven by interview data as the test takers report to have understood OE text better and it required more focused reading. The investigation of processes the test takers went through on an item basis shows that MC vocabulary questions are answered by background knowledge most of the time rather than by reading the context and guessing from the context as it is intended. In this context, the validity of vocabulary questions in MC format is questionable and they need to be examined in a more detailed way to assure whether they really trigger the intended processes or not. The systematic comparison of eye movement and retrospective verbal reports shed light into the processes of test takers and revealed how they managed the task. The study showed that OE tends to trigger higher cognitive processes when compared to MC.

The study also showed that the triangulation of eye-tracking technology and retrospective verbal reports can be very promising in revealing cognitive processes and the fact that all of these data sets in the study yielded harmonious results suggests that they should be used in combination.

The study has several limitations. First of all, the number of statistical analysis in eye-movement data was limited in the study as the study mostly attempts to have a descriptive nature. Secondly, the participant group was homogenous in terms of their high language level and for this reason, the use of construct-irrelevant strategy use was very limited. Conducting the same study with a less homogeneous group including lower level language users could yield more various strategies and more interesting results. The data were collected in a lab environment, not in a real classroom environment and this might have had an effect on the scores. Next, the chin rests were used in the study and that might have affected the performance of test takers although all participants claimed that it did not. Lastly, only the texts were counterbalanced in two conditions and the lack of counterbalancing in test methods was another limitation of the study.

The current study has some suggestions to the stakeholders. First of all, although the practicality of its scoring makes multiple-choice items very popular and they are widely used, it should be noted that they can disrupt reading processes and if possible, open-ended items should be preferred. The fact that multiple-choice questions lead test takers to come up with a number of test-taking strategies to cope with the item demands raises concerns about the validity of these item types. It is also known that options are prepared based on the comprehension of the test writer and most of the time, they are prepared with subtle differences among them which can be tricky and misleading for test takers. As expected, test takers also devise some

ways to eliminate these options in a strategic way. This might even cause them to engage in construct-irrelevant strategies through which they use some peripheral ways to answer the questions without undergoing intended reading processes. When the worldwide importance of tests is considered as “gatekeepers to the institutions” (Shohamy, 2001), the seriousness of the problem can be realized. As testing and assessment can be regarded as a chain of decision-making, a faulty decision at any part of this chain can result in serious problems for many stakeholders. Using a test method merely because of its practicality by disregarding all the concerns raised about its validity may mean a crucial mistake is made in making an important decision on test takers. Therefore, so as to make valid decisions based on the scores, first, the test should be proven to be valid. In addition, item writers are required to make sure that the items they write are also cognitively valid by piloting through several methods such as the ones used in this study to make sure whether the items they have written trigger any construct-irrelevant strategies or irrelevant processes. They should revise their items accordingly. Test takers should be also careful not to depend on the test-management strategies too much as the same strategies can also cause them to answer an item incorrectly. Most importantly, MC reading exercises should be used cautiously in teaching reading as they do not reflect authentic reading skills.

APPENDIX A

CONTEXTUAL PARAMETERS: ITEM-LEVEL

	Content dimension	1_1	a. main idea	b. detail	1_2	a. fact	b. opinion	
I1	Explicitness dimension	1_3	a. from explicit information		b. from implicit information			
	Where did you find the information to answer the question?	1_4	a. within sentence		b.across sentences		c. at the whole text level	
	Ease	1_5	a. easy	b.moderate		c.difficult	d. too difficult	
	Content dimension	2_1	a. main idea	b. detail	2_2	a. fact	b. opinion	
I2	Explicitness dimension	2_3	a. from explicit information		b. from implicit information			
	Where did you find the information to answer the question?	2_4	a. within sentence		b.across sentences		c. at the whole text level	
	Ease	2_5	a. easy	b.moderate		c difficult	d. too difficult	
	Content dimension	3_1	a. main idea	b. detail	3_2	a. fact	b. opinion	
I3	Explicitness dimension	3_3	a. from explicit information		b. from implicit information			
	Where did you find the information to answer the question?	3_4	a. within sentence		b.across sentences		c. at the whole text level	
	Ease	3_5	a. easy	b.moderate		c.difficult	d. too difficult	

APPENDIX B

CONTEXTUAL PARAMETERS: TEXT -LEVEL

Domain		1. social		2. professional		3. academic
Discourse	Genre	1. public sign/notice	2. magazine /newspaper article/report	3. advertisement/leaflet /brochure	4. letter/memo /email message	5.fiction book/ excerpt
Mode	Tone	1. narrative	2. expository	3. argumentative	4. instructive	5. informative
Rhetorical organization		To what extent are the test takers familiar with the organizational structure of the text?				
		1. familiar	2.	3.	4.	5. not familiar
Subject specificity		Is the topic of the text of general interest so that it can appeal to the majority of test takers?				
		1. general	2.	3.	4.	5. specific
Cultural specificity		Is the topic of the text culture-neutral or is it loaded with specific cultural content?				
		1. culture neutral	2. somehow culture specific	5. culture specific		
Text abstractness		To what extent is the text concrete/abstract?				
		1. concrete	2.	3.	4.	5. abstract
Text ease/Readability		Is the text manageable in terms of comprehension on part of the test takers?				
		1. manageable	2.	3.	4.	5. not manageable
Level (CEFR)		1. B2-	2. B2	3. B2+	4.C1	

APPENDIX C

TEXT 1 MULTIPLE-CHOICE FORMAT

THE ORIGINS OF CETACEANS

1 It should be obvious that cetaceans—whales and dolphins—are mammals. They breathe through lungs and give birth to live young. Their streamlined bodies, the absence of hind legs, and the presence of a fluke[1] and blowhole[2] cannot disguise their relationship with mammals living on land. However, unlike the cases of other animals such as seals and sea lions whose limbs are functional both on land and at sea), it is not easy to visualize what the first whales looked like. Extinct but already fully marine cetaceans are known from the fossil record. How was the gap between a walking mammal and a swimming whale bridged? Intermediate fossils between land mammals and cetaceans were missing until recently.

2 Very exciting discoveries have finally allowed scientists to reconstruct the most likely origins of cetaceans. In 1979, a team looking for fossils in northern Pakistan found what proved to be the oldest fossil whale. The fossil was officially named Pakicetus. Pakicetus was found embedded in rocks formed from river deposits that were fifty-two million years old. The river was actually not far from an ancient ocean known as the Tethys Sea.

3 The Pakicetus fossil consists of a complete skull of an extinct group of ancestors of modern cetaceans. Although limited to a skull, the Pakicetus fossil provides precious details on the origins of cetaceans. The skull shape is cetacean-like but its jawbones lack the enlarged space that is filled with fat or oil and used for receiving underwater sound in modern whales. Pakicetus probably detected sound through the ear opening as in land mammals. The skull also lacks a blowhole, another cetacean adaptation for diving. Other features, however, show experts that Pakicetus is a transitional form between a group of extinct flesh-eating mammals and cetaceans. It has been suggested that Pakicetus fed on fish in shallow water and was not yet adapted for life in the open ocean. It probably bred and gave birth on land.

4 Another major discovery was made in Egypt in 1989. Several skeletons of another early whale, Basilosaurus, were found in sediments left by the Tethys Sea and now **exposed** in the Sahara desert. This whale lived twelve million years after Pakicetus. Many incomplete skeletons were found but they included, for the first time a complete hind leg that features a foot with three tiny toes. Such legs would have been far too small to have supported such a big animal on land. Basilosaurus was undoubtedly a fully marine whale with possibly nonfunctional hind legs.

5 An even more exciting find was reported in 1994, also from Pakistan. The now extinct whale Ambulocetus natans (the walking whale that swam) lived around three million years after Pakicetus but nine million years before Basilosaurus. The fossil luckily includes a good portion of the hind legs. The presence of hind legs clearly bridged the gap between a walking mammal and swimming whale. The legs were strong and ended in long feet very much like those of a modern seal. The legs were certainly functional both on land and at sea. This clearly proved Ambulocetus to be a transition species. Moreover, the whale retained a tail and lacked a fluke, the major means of moving in modern cetaceans. The structure of the backbone shows, however, that Ambulocetus swam like modern whales by moving the rear portion of its body up and down. The large hind legs were used for moving forward in water. On land, where it probably bred and gave birth, Ambulocetus may have moved around very much like a modern sea lion. It was undoubtedly a whale that linked life on land with life at sea.

1. Fluke: the two parts that constitute the large triangular tail of a whale
2. Blowhole: a hole in the top of the head used for breathing.

1.In paragraph 1, what does the author say about the presence of a blowhole in cetaceans?

a) It clearly indicates that cetaceans are mammals.
b) It cannot conceal the fact that cetaceans are mammals.
c) It is the main difference between cetaceans and land-dwelling mammals.
d) It cannot yield clues about the origins of cetaceans.

2. Pakicetus and modern cetaceans have similar _____ .

a) hearing structures
b) adaptations for diving
c) skull shapes
d) breeding locations

3. The word "exposed" in the passage is closest in meaning to ___ .

a) explained b) visible c) identified d) located

4. The hind leg of Basilosaurus was a significant find because it showed that Basilosaurus _____ .

a) lived later than Ambulocetus Natans
b) lived at the same time as Pakicetus
c) was able to swim well
d) could not have walked on land

5. It can be inferred that Basilosaurus bred and gave birth in which of the following locations?

a) On land
b) Both on land and at sea
c) In shallow water
d) In a marine environment

6. Why does the author use the word "luckily" in mentioning that Ambulocetus natans fossil included hind legs?

a) Fossil legs of early whales are a rare find.
b) The legs provided important information about the evolution of cetaceans.
c) The discovery allowed the scientists to reconstruct a complete skeleton of the whale.
d) Until that time, only the front legs of early whales had been discovered.

APPENDIX D

TEXT 2 MULTIPLE-CHOICE FORMAT

SWIMMING MACHINES

1 Tunas, mackerels, and billfishes (marlins, sailfishes, and swordfish) swim continuously. Feeding, courtship, reproduction, and even "rest" are carried out while in constant motion. As a result, practically every aspect of the body form and function of these swimming "machines" is adapted to enhance their ability to swim.

2 Many of the adaptations of these fishes serve to reduce water resistance. Interestingly enough, several of these hydrodynamic adaptations resemble features designed to improve aerodynamics in engineering. Though human engineers are new to the game, tunas and their relatives evolved their "high-tech" designs long ago.

3 Tunas, mackerels, and billfishes have made streamlining into an art form. Their bodies are sleek and compact. The body shapes of tunas, in fact, are nearly ideal from an engineering point of view. Most species lack scales over most of the body. This feature makes their bodies smooth and slippery. They also have a slick and transparent cover that reduces water resistance. The fins are stiff, smooth, and narrow. These qualities also help reduce water resistance. When not in use, the fins are tucked into special grooves or depressions so that they lie at the same level with the body. Therefore, they do not break up its smooth contours. Airplanes retract their landing gear while in flight for the same reason.

4 Tunas, mackerels, and billfishes have even more sophisticated adaptations than these to improve their hydrodynamics. The long bill of marlins, sailfishes, and swordfish probably helps them move smoothly the water. Many supersonic aircraft have a similar needle at the nose. Most tunas and billfishes have a series of keels and finlets near the tail. Although most of their scales have been lost, tunas and mackerels retain a patch of coarse scales called the corselet. The keels, finlets, and corselet help direct the flow of water over the body surface to reduce water resistance. Again, supersonic jets have similar features.

5 Because they are always swimming, tunas simply have to open their mouths and water is forced in and over their gills. Accordingly, they have lost most of the muscles that other fishes use to suck in water and push it past the gills. In fact, tunas must swim to breathe. They must also keep swimming to keep from sinking, since most have largely or completely lost the swim bladder. Swim bladder helps most other fish remain floating.

6 One potential problem is that opening the mouth to breathe detracts from the streamlining of these fishes and tends to slow them down. Some species of tuna have specialized grooves in their tongue. It is thought that these grooves help to ***channel*** water through the mouth and out the gill slits. This process reduces the water resistance.

7 There are adaptations that increase the amount of forward thrust as well as those that reduce water resistance among fast swimming fishes. Perhaps most important of all is their ability to make use of swirls and eddies (circular currents) in the water. They can glide past eddies that would slow them down and then gain extra thrust by "pushing off" the eddies. Scientists and engineers are beginning to study this ability of fishes in the hope of designing more efficient propulsion systems for ships.

8 These fishes also have a highly efficient mechanism that maintains a warm body temperature. With the help of this mechanism, a bluefin tuna in the water of seven°C (fortyfive°F) can maintain a core temperature of over twenty-five °C (seventy-seven°F). This warm body temperature may help not only the muscles to work better, but also the brain and the eyes. The billfishes have gone one step further. They have evolved special "heaters" of modified muscle tissue that warm the eyes and brain, maintaining peak performance of these critical organs.

1. Why does the author mention that "*Airplanes retract their landing gear while in flight*?"

a) To show air resistance and water resistance work differently from each other.
b) To argue that some fishes are better designed than airplanes are.
c) To provide evidence that airplane engineers have studied the design of fish bodies.
d) To demonstrate a similarity in design between certain fishes and airplanes.

2. According to paragraph 4, the long bills of marlins, sailfish, and swordfish probably help these fishes by ____ .

a) increasing their ability to defend themselves
b) allowing them to change direction easily
c) increasing their ability to detect odors
d) reducing water resistance as they swim

3. According to the passage, which of the following is one of the reasons that tunas are in constant motion?

a) They lack a swim bladder.
b) They need to suck in more water than other fishes do.
c) They have large muscles for breathing.
d) They cannot open their mouth unless they are in motion.

4. The word "channel" in the passage is closest in meaning to ___ .

a) reduce b) remove c) direct d) provide

5. According to the passage, one of the adaptations of fast-swimming fishes that might be used to improve the performance of ships is these fishes' ability to ____ .

a) swim directly through eddies
b) make efficient use of water currents
c) cover great distances without stopping
d) gain speed by forcing water past their gills

6. According to paragraph 8, which of the following is true of bluefin tunas?

a) Their eyes and brain are more efficient than those of any other fish.
b) Their body temperature can change greatly depending on the water temperature.
c) They can swim in waters that are much colder than their own bodies.
d) They have special muscle tissue that warms their eyes and brain.

APPENDIX E

TEXT 1 OPEN-ENDED FORMAT

THE ORIGINS OF CETACEANS

1 It should be obvious that cetaceans—whales and dolphins—are mammals. They breathe through lungs and give birth to live young. Their streamlined bodies, the absence of hind legs, and the presence of a fluke[1] and blowhole[2] cannot disguise their relationship with mammals living on land. However, unlike the cases of other animals such as seals and sea lions whose limbs are functional both on land and at sea), it is not easy to visualize what the first whales looked like. Extinct but already fully marine cetaceans are known from the fossil record. How was the gap between a walking mammal and a swimming whale bridged? Intermediate fossils between land mammals and cetaceans were missing until recently.

2 Very exciting discoveries have finally allowed scientists to reconstruct the most likely origins of cetaceans. In 1979, a team looking for fossils in northern Pakistan found what proved to be the oldest fossil whale. The fossil was officially named Pakicetus. Pakicetus was found embedded in rocks formed from river deposits that were fifty-two million years old. The river was actually not far from an ancient ocean known as the Tethys Sea.

3 The Pakicetus fossil consists of a complete skull of an extinct group of ancestors of modern cetaceans. Although limited to a skull, the Pakicetus fossil provides precious details on the origins of cetaceans. The skull shape is cetacean-like but its jawbones lack the enlarged space that is filled with fat or oil and used for receiving underwater sound in modern whales. Pakicetus probably detected sound through the ear opening as in land mammals. The skull also lacks a blowhole, another cetacean adaptation for diving. Other features, however, show experts that Pakicetus is a transitional form between a group of extinct flesh-eating mammals and cetaceans. It has been suggested that Pakicetus fed on fish in shallow water and was not yet adapted for life in the open ocean. It probably bred and gave birth on land.

4 Another major discovery was made in Egypt in 1989. Several skeletons of another early whale, Basilosaurus, were found in sediments left by the Tethys Sea and now exposed in the Sahara desert. This whale lived twelve million years after Pakicetus. Many incomplete skeletons were found but they included, for the first time a complete hind leg that features a foot with three tiny toes. Such legs would have been far too small to have supported such a big animal on land. Basilosaurus was undoubtedly a fully marine whale with possibly nonfunctional hind legs.

5 An even more exciting find was reported in 1994, also from Pakistan. The now extinct whale Ambulocetus natans (the walking whale that swam) lived around three million years after Pakicetus but nine million years before Basilosaurus. The fossil luckily includes a good portion of the hind legs. The presence of hind legs clearly bridged the gap between a walking mammal and swimming whale. The legs were strong and ended in long feet very much like those of a modern seal. The legs were certainly functional both on land and at sea. This clearly proved Ambulocetus to be a transition species. Moreover, the whale retained a tail and lacked a fluke, the major means of moving in modern cetaceans. The structure of the backbone shows, however, that Ambulocetus swam like modern whales by moving the rear portion of its body up and down. The large hind legs were used for moving forward in water. On land, where it probably bred and gave birth, Ambulocetus may have moved around very much like a modern sea lion. It was undoubtedly a whale that linked life on land with life at sea.

1. Fluke: the two parts that constitute the large triangular tail of a whale
2. Blowhole: a hole in the top of the head used for breathing.

1. In paragraph 1, what can the presence of a blowhole in cetaceans not conceal according to the author?

_____

2. What do Pakicetus and modern cetaceans have in common?

_____

3. Find and write a word which means "visible" in paragraph 4.

_____

4. The hind legs of Basilosaurus show that it couldn't have walked on foot because

_____

5. Find and write a sentence in paragraph 4 that supports the inference that "Basilosaurus bred and gave birth in water".

_____

6. Why does the author use word "luckily" while mentioning that Ambulocetus natans fossil included hind legs?

_____

APPENDIX F

TEXT 2 OPEN-ENDED FORMAT

SWIMMING MACHINES

1 Tunas, mackerels, and billfishes (marlins, sailfishes, and swordfish) swim continuously. Feeding, courtship, reproduction, and even "rest" are carried out while in constant motion. As a result, practically every aspect of the body form and function of these swimming "machines" is adapted to enhance their ability to swim.

2 Many of the adaptations of these fishes serve to reduce water resistance. Interestingly enough, several of these hydrodynamic adaptations resemble features designed to improve aerodynamics in engineering. Though human engineers are new to the game, tunas and their relatives evolved their "high-tech" designs long ago.

3 Tunas, mackerels, and billfishes have made streamlining into an art form. Their bodies are sleek and compact. The body shapes of tunas, in fact, are nearly ideal from an engineering point of view. Most species lack scales over most of the body. This feature makes their bodies smooth and slippery. They also have a slick and transparent cover that reduces water resistance. The fins are stiff, smooth, and narrow. These qualities also help reduce water resistance. When not in use, the fins are tucked into special grooves or depressions so that they lie at the same level with the body. Therefore, they do not break up its smooth contours. Airplanes retract their landing gear while in flight for the same reason.

4 Tunas, mackerels, and billfishes have even more sophisticated adaptations than these to improve their hydrodynamics. The long bill of marlins, sailfishes, and swordfish probably helps them move smoothly the water. Many supersonic aircraft have a similar needle at the nose. Most tunas and billfishes have a series of keels and finlets near the tail. Although most of their scales have been lost, tunas and mackerels retain a patch of coarse scales called the corselet. The keels, finlets, and corselet help direct the flow of water over the body surface to reduce water resistance. Again, supersonic jets have similar features.

5 Because they are always swimming, tunas simply have to open their mouths and water is forced in and over their gills. Accordingly, they have lost most of the muscles that other fishes use to suck in water and push it past the gills. In fact, tunas must swim to breathe. They must also keep swimming to keep from sinking, since most have largely or completely lost the swim bladder. Swim bladder helps most other fish remain floating.

6 One potential problem is that opening the mouth to breathe detracts from the streamlining of these fishes and tends to slow them down. Some species of tuna have specialized grooves in their tongue. It is thought that these grooves help to channel water through the mouth and out the gill slits. This process reduces the water resistance.

7 There are adaptations that increase the amount of forward thrust as well as those that reduce water resistance among fast swimming fishes. Perhaps most important of all is their ability to make use of swirls and eddies (circular currents) in the water. They can glide past eddies that would slow them down and then gain extra thrust by "pushing off" the eddies. Scientists and engineers are beginning to study this ability of fishes in the hope of designing more efficient propulsion systems for ships.

8 These fishes also have a highly efficient mechanism that maintains a warm body temperature. With the help of this mechanism, a bluefin tuna in the water of seven°C (fortyfive°F) can maintain a core temperature of over twenty-five°C (seventy-seven°F). This warm body temperature may help not only the muscles to work better, but also the brain and the eyes. The billfishes have gone one step further. They have evolved special "heaters" of modified muscle tissue that warm the eyes and brain, maintaining peak performance of these critical organs.

1. The writer mentions that the design of some fish bodies is used by engineers. Find the example about this similarity in paragraph 3.

_____.

2. It can be understood that the adaptations such as fins and long bill of some types of fishes help them to swim smoothly by _____

_____.

3. According to the text, tunas have gone through major evolutionary changes. What are two major reasons why tunas must be in constant motion?

_____

_____.

4.Find and write a word which means "to direct" in paragraph 6?

_____.

5. According to the passage, one of the adaptations of fast swimming fishes that can also be used to improve the performance of ships is their ability to _____

_____.

6. What makes bluefin tunas swim in waters that are much colder than their bodies?

_____.

APPENDIX G

INFORMED CONSENT FORM

KATILIMCI BİLGİ ve ONAM FORMU

Araştırmayı destekleyen kurum: Boğaziçi Üniversitesi

Araştırmanın adı: A qualitative exploration of reading and test-taking strategy use depending on test-format: multiple-choice vs. open-ended questions- Test formatına bağlı olarak- çoktan seçmeli ve açık uçlu sorularda kullanılan okuma ve test çözme stratejilerinin niteliksel bir araştırması

Proje Yürütücüsü: Aylin Ünaldı (Boğaziçi Üniversitesi, Eğitim Fakültesi)

Kurumsal adres: Boğaziçi Üniversitesi, Eğitim Fakültesi, Yabancı Diller Eğitimi Bölümü, 34342 Bebek İstanbul Telefon: 0212 3594609 e-posta: aunaldi@boun.edu.tr

Araştırmacının adı: Hatice Akgün

E-mail adresi: haticevcil@hotmail.com

Telefonu: 506 703 65 67

Proje konusu: Okuma testlerini çözerken öğrencilerin kullandıkları okuma ve test çözme stratejileri testin geçerliliğini belirleme konusunda önemli bir rol oynar. Kullanılan stratejiler ise testin soru formatına bağlı olarak değişiklik gösterebilir. Testte kullanılan stratejilerin, testin ölçmeyi hedeflediği bilişsel süreçlerle örtüşmesi beklenir. Bu çalışma, ileri düzeyde İngilizce okuma sınavını çözerken, öğrencilerin çoktan seçmeli veya açık uçlu soru formatlarına bağlı olarak kullanmış oldukları okuma ve sınav çözme stratejilerinde meydana gelen farklılıkları ortaya çıkarmayı amaçlayan bir araştırmadır. Sınav çözerken kullanılan bilişsel süreçleri öğrenmek için öğrencilerin aynı okuma parçasını iki farklı test formatında çözerken göz hareketleri kaydedilecek ve daha sonrasında da öğrencilerden sınavı nasıl çözdüklerini anlatmaları istenilecek. Araştırmanın sonuçları hangi soru formatının okuma becerisini ölçmede daha geçerli olduğunu gösterecek.

Onam: Bu çalışmanın sonucunda, her bir test formatında kullanılan okuma ve test çözme stratejilerine bağlı olarak, hangi soru formatının okuma becerisini ölçmede daha etkili olduğunu öğrenmeyi hedefliyoruz.

Araştırmaya katılmayı kabul ettiğiniz takdirde, Boğaziçi Üniversitesi Güney Kampüste Psikoloji bölümüne ait görme laboratuvarında yaklaşık olarak 2 saat sürecek bir

çalışmaya bireysel olarak katılacaksınız. Bu çalışma kapsamında, biri açık uçlu ve biri çoktan seçmeli olmak üzere toplam 2 okuma sınavı çözeceksiniz. Bu sınavları göz hareketlerinizin kaydedilmesi için bilgisayar ekranından çözeceksiniz ve sonra da sınavı nasıl çözdüğünüzü araştırmacıya anlatacaksınız. Araştırmacı size bu konuda yardımcı olmak için sizi yönlendirecek sorular soracak ve veriyi daha sonra analiz edebilmek için konuşmanızı kayıt altına alacak.

Bu çalışma boyunca, isminiz ve ses kayıtları tamamen gizli tutulacaktır ve çalışmanın herhangi bir aşamasında isminiz kullanılmayacaktır.

Çalışmaya katılmanız tamamen isteğe bağlıdır. Bu çalışmaya katılıp katılmamanız, ders notlarınızı hiçbir şekilde- olumlu ya da olumsuz- etkilemeyecektir. Sizden ücret talep etmiyoruz ancak ayıracağınız vakit bilimsel bir çalışmanın gerçekleştirilmesine katkıda bulunacaktır. Ayrıca çalışmaya katılmanızın karşılığında size USB Bellek ya da kulaklık hediye edilecektir.

Sizden alınan veriler ileride başka çalışmalar için de kullanılabilir. Katıldığınız takdirde çalışmanın herhangi bir aşamasında herhangi bir sebep göstermeden onayınızı çekmek hakkına da sahipsiniz.

Yapılacak çalışmanın öngörülen hiçbir olumsuz etkisi ya da ayıracağınız 2 saat dışında size hiçbir yükü olmayacaktır. Çalışma sırasında yapılan analiz sonuçları isterseniz kişisel olarak sizinle paylaşılacaktır.

Çalışma sırasında ve sonucunda ortaya çıkan bilgiler bizi açık uçlu soruların mı, çoktan seçmeli soruların mı okuma becerisini ölçmede daha etkili olduğu konusunda ve ayrıca hangi soru formatında hangi okuma stratejilerinin kullanıldığı konusunda aydınlatacaktır.

Bu bilgiler, size de kişisel olarak ileri düzeyde okuma becerinizi ve strateji kullanımını geliştirmede faydalı olabilir.

Bu formu imzalamadan önce, çalışmayla ilgili sorularınız varsa lütfen sorun. Daha sonra sorunuz olursa, Hatice Akgün'e (Telefon:0506 703 65 67) sorabilirsiniz. Araştırmayla ilgili haklarınız konusunda Boğaziçi Üniversitesi İnsan Araştırmaları Etik Alt Kurulu (INAREK) veya INAREK/SBB Etik Alt Kurulu kurullarına da danışabilirsiniz.

Bana anlatılanları ve yukarıda yazılanları anladım. Bu formun bir kopyasını aldım. Çalışmaya katılmayı kabul ediyorum.

Katılımcı Adı-Soyadı:.....

İmzası: .....

Tarih (gün/ay/yıl):..... /..... /,

APPENDIX H

AREAS OF INTEREST

Table H1. Areas of Interest for Text 1

No	Relevant paragraph	Key words/phrases/ sentences
1	Paragraph 1	Key 1: Lines 2-3
2	Paragraph 3	Key 1: Lines 2-3
3	Paragraph 4	Key 1: Lines 1&2
4	Paragraph 4	Key 1: Lines 3&4
5	Paragraph 4	Key1: Lines 3&4&5 –
6	Paragraph 5 (&paragraph 1)	Key 1: Lines 3,4,5,6 Key 2: Paragraph 5 – whole paragraph Key 3: Paragraph 1 – How was the gap between walking mammal and swimming whale bridged? Intermediate fossils were missing until recently.

Table H2. Areas of Interest for Text 2

No	Relevant paragraph	Key words/phrases/ sentences
1	Paragraph 3	Key 1: Lines 5,6,7
2	Paragraphs 2,3,4	Key 1: Paragraph 2, line 1: Many of the adaptations of these fishes serve to reduce water resistance Key 2: Paragraph 3, line 4: The fins are stiff, smooth and narrow. These qualities also help reduce water resistance. (ONLY FOR FORM D) Key 3: Paragraph 4, line 1-2
3	Paragraph 5	Key 1: lines 3&4
4	Paragraph 6	Key 1: Lines 2 &3
5	Paragraph 7	Key 1: line 2,3,4,5,
6	Paragraph 8	Key 1: line 1,2,3

APPENDIX I

READING STRATEGIES CODING RUBRIC

Strategy	*Description*
	Prior to test taking
R1	reads the text first carefully before attempting the task
R2	reads the text expeditiously to have a general idea before attempting the task
	Expeditious Reading
R3	Rapidly looks for and matches figures, dates, names, words etc in the text and question.
R4	Looks for markers in the text such as connectors, grammatical structures, examples to locate information.
R5	Tries to understand the information in the text quickly through skimming.
R6	Searches for key words / ideas in the text related to the general topic of the question
	Other Reading Strategies
R8	Reads only the part of the text which seems related to specific questions.
	Careful Reading
R9	Focuses on the parts of a sentence to understand it clearly
R10	Reads carefully across sentences (to establish the connections of ideas between sentences or parts of the text by identifying relationships such cause and effect, claim and supports etc.)
R11	Tries to create a textual representation by establishing between paragraphs to understand the organisation of information in the whole text
R14	Rereads the important or difficult parts of the text
R7	Makes inferences based on the information in the text.

R1- R8, and R14 are adapted from Cohen, A. D., & Upton, T. A. (2006). Strategies in responding to the new TOEFL reading tasks. ETS Research Report Series, 2006(1).

R7, R9, R10, are the patterns that emerged from the data and they were developed accordingly.

APPENDIX J

TEST- MANAGEMENT STRATEGIES CODING RUBRIC

Strategies	*Description*
T1	Goes back to the question for clarification: Rereads, translates the question or wrestles with the question intent.
T4	Reads the question and the options to form an idea before going back to text.
T5	Skips the difficult question and moves on to the next question. Then, comes back to the skipped question.
T6	Uses the order of questions in the text as a clue to locate the text span where the answer is located.
T8	Predicts or produces own answer after reading the portion of the text referred to by the question.
T11	Considers the options and identifies an option with an unknown vocabulary.
T14	Considers the options and selects preliminary option(s) (lack of certainty indicated).
T17	Considers the options and eliminates the ones that are similar or overlapping.
T18	Considers the options and wrestles with the option meaning
T19	Makes an educated guess (e.g., using background knowledge or extra-textual knowledge).
T22	Selects options through elimination of other option(s) as unreasonable based on background knowledge.
T23	Selects options through elimination of other option(s) as unreasonable based on paragraph/overall passage meaning.
T29	Expresses uncertainty at correctness of an answer chosen
T30	Stops reading the options when they reach an answer
T31	Receives meaning clues from answering one question that are helpful in answering another
T35	Continues to read the relevant paragraph when the answer is found. (to make sure the response is correct)
T36	Stops reading the relevant paragraph when the answer is found.

Strategies T1, T4 & T8, T11, T14, T18, T19, T22, T23 are taken from 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), 209-250. doi:10.1177/0265532207076364

Strategies T6, T10 are taken from Ünaldı, A. 2004. Construct validation of the reading subskills of the Boğaziçi University English Proficiency Test. Unpublished PhD Thesis. Faculty of Education: Boğaziçi University.

Strategy T5 is taken from Lim, H. J. (2014). Exploring the validity evidence of the TOEFL iBT reading test from a cognitive perspective

Strategies T29, T30, T31 are taken from Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. Language Testing, 8(1), 41-66.

APPENDIX K

TEST-WISENESS STRATEGIES CODING RUBRIC

Strategy	*Description*
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 passage in it—possibly a key word.
TW4	Chooses a word / phrase as an answer when it is in the same sentence with the key words.

Strategies TW1, TW2, TW3 are taken from 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), 209-250. doi:10.1177/0265532207076364

TW4 emerged from the data.

APPENDIX L

EYE-TRACKING STATISTICS FOR CORRECT RESPONSES

Table L1. Eye-Tracking Statistics for Item 1

		QS TFC	QS TRT	AoI TFC	AoI TRT
MC (N=31)	Mean	2.23	2.10	4.60	4.09
	St.Dev.	0.78	1.06	2.36	2.09
OE (N=31)	Mean	2.26	1.97	9.67	10.49
	St.Dev	1.17	1.04	4.26	3.93
Mann Whitney U		92.00	85.00	24.00	14.00
Z		0.00	-0.32	-3.07	-3.52
Sig (2-tailed)		100.00	.774	.001*	.000*

Table L2. Eye-Tracking Statistics for Item 2

		QS TFC	QS TRT	AoI TFC	AoI TRT
MC (N=31)	Mean	2.03	1.99	6.17	6.17
	St.Dev.	1.00	1.28	2.23	2.36
OE (N=31)	Mean	2.61	2.58	9.38	9.97
	St.Dev	1.82	2.07	3.71	4.67
Mann Whitney U		160.00	161.00	91.00	95.00
Z		-0.69	-0.66	-2.66	-2.54
Sig (2-tailed)		.507	.525	.007*	.010*

Table L3. Eye-Tracking Statistics for Item 3

		QS TFC	QS TRT	AoI TFC	AoI TRT
MC (N=31)	Mean	1.52	1.46	3.98	4.15
	St. Dev.	0.58	0.64	1.81	2.05
OE (N=31)	Mean	2.75	2.38	7.36	8.56
	St.Dev	1.31	1.29	3.81	4.22
Mann Whitney U		80.00	108.0	78.00	58.00
Z		-3.10	-2.35	-3.15	-3.68
Sig (2-tailed)		.001*	.018*	.001*	.000*

Table L4. Eye-Tracking Statistics for Item 4

		QS TFC	QS TRT	AoI TFC	AoI TRT
MC (N=31)	Mean	2.59	2.90	6.63	6.80
	St. Dev.	1.17	1.12	3.26	3.47
OE (N=31)	Mean	2.14	2.09	10.16	11.82
	St. Dev	1.13	1.15	8.00	9.18
Mann Whitney U		176.00	135.00	190.00	162.00
Z		-0.86	-1.94	-0.50	-1.23
Sig (2-tailed)		0.39	0.05	0.62	0.22

Table L5. Eye-Tracking Statistics for Item 5

		QS TFC	QS TRT	AoI TFC	AoI TRT
MC (N=31)	Mean	2.00	1.97	7.73	7.89
	St. Dev.	1.32	1.39	2.49	2.79
OE (N=31)	Mean	2.93	2.71	7.16	7.67
	St. Dev.	1.54	1.58	3.99	4.78
Mann Whitney U		183.00	206.00	233.00	250.00
Z		-2.18	-1.71	-1.16	-0.81
Sig (2-tailed)		0.03	0.09	0.25	0.42

Table L6. Eye-Tracking Statistics for Item 6

		QS TFC	QS TRT	AoI TFC	AoI TRT
MC (N=31)	Mean	1.68	1.79	7.81	7.71
	St. Dev.	0.86	1.22	4.20	4.06
OE (N=31)	Mean	1.60	1.46	5.06	5.05
	St. Dev.	0.77	0.83	3.86	3.84
Mann Whitney U		330.00	299.00	189.00	193.00
Z		-0.11	-0.68	-2.70	-2.63
Sig (2-tailed)		0.91	0.50	0.01	0.01

APPENDIX M

RESULTS OF T-TEST FOR READING STRATEGIES

Table M1. Descriptive Statistics for Reading Strategies

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	MC	65,27	11	77,183	23,272
	OE	80,09	11	86,103	25,961

Table M2. T-Test Results for Reading Strategies

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Dev.	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	MC OE	-14.818	14.865	4.482	-24.805	-4.832	-3.306	10	.008

APPENDIX N

RESULTS OF T-TEST FOR TEST-TAKING STRATEGIES

Table N1. Descriptive Statistics for Test-Taking Strategies

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	MC	21.10	21	20.964	4.575
	OE	11.38	21	17.985	3.925

Table N2. T-Test Results for Test-Taking Strategies

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Dev.	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	MC-OE	9.714	20.616	4.499	.330	19.099	2.159	20	.043

REFERENCES

Afflerbach, P., Pearson, P. D., & Paris, S. G. (2008). Clarifying differences between reading skills and reading strategies. *The Reading Teacher*, *61*(5), 364-373. doi:10.1598/RT.61.5.1.

Alderson, J. C. (2000). *Assessing reading.* Cambridge, UK; New York, NY, USA: Cambridge University Press.

Allan, A. (1992). Development and validation of a scale to measure test- wiseness in EFL/ESL reading test takers. *Language Testing*, *9*(2), 101-119. doi:10.1177/026553229200900201.

Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing*, *8*(1), 41-66. doi: 10.1177/026553229100800104

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

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

Bax, S., & Weir, C. (2012). Investigating learners' cognitive reading processes during a computer-based CAE reading test. *Research Notes*, *47*, 3–14. Retrieved from www.cambridgeesol.org/rs_notes/rs_nts47.pdf.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye tracking. *Language Testing*, *30*(3), 1–25. doi:10.1177/0265532212473244.

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

Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics, 25,* 133-150. doi: 10.1017/S0267190505000073.

Bernhardt, E., & Kamil, M. L. (1995). Interpreting relationships between L1 and L2 reading: Consolidating the linguistic threshold and the linguistic interdependence hypotheses. *Applied Linguistics*, *16*(1), 15-34.

Bowles, M. A. (2010). *The think-aloud controversy in second language research*. New York: Routledge.

Brysbaert, M., and Vitu, F. (1998). Word skipping: implications for theories of eye movement control in reading. In G. Underwood (ed.), *Eye guidance in reading and scene perception* (pp. 125–48). Oxford: Elsevier.

Carpenter, P. A., and Just, M. A. (1983). What your eyes do while your mind is reading. In K. Rayner (ed.), *Eye movements in reading: Perceptual and language processes* (pp. 275–307). New York: Academic Press.

Chamot, A. & O'Malley, J. (1994). Instructional approaches and teaching procedures. In R. Spangenberg-Urbschat & R. Pritchard (Eds), *Kids come in all languages: Reading instruction for ESL students* (pp. 82-107). Delaware, USA: International Reading Association (IRA).

Cohen, A. D. (1998). *Strategies in learning and using a second language*. London: Longman.

Cohen, A. D. and Upton, T. A. (2006). *Strategies in responding to the new TOEFL reading tasks* (TOEFL Monograph Series Report No. 33). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-06-06.pdf.

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), 209–250. doi:10.1177/0265532207076364.

Cohen, A. D. (2012a). Test-taking strategies. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (eds.), *The Cambridge guide to assessment* (pp. 96-104). Cambridge: Cambridge University Press.

Cohen, A. D. (2012b). Test taker strategies and task design. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing in a nutshell* (pp. 262-277). Abingdon, England: Routledge.

Cohen, A. D. (2013). Using test-wiseness strategy research in task development. In A. J. Kunnan (Ed.), *The companion to language assessment – Vol. 2: Approaches and development, Part 7: Assessment development*. Hoboken, NJ: Wiley and Sons.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302. doi: 10.1037/h0040957.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, *49*(2), 222-251. doi: 10.3102/00346543049002222.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis: Verbal reports as data.* London: The MIT Press.

Field, J. (2011). Into the mind of the academic listener. *Journal of English for Specific Purposes,10(2),* 102-112. doi: 10.1016/j.jeap.2011.04.002.

Field, J. (2012). The cognitive validity of the lecture-based question in the IELTS listening paper. In P. Thompson (Ed.), *IELTS Research Reports* (Vol. 9) (pp. 17–66). London: British Council.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London: Routledge.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: L. Erlbaum Associates.

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

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

Green, A., & University of Cambridge. Local Examinations Syndicate. (1998). *Verbal protocol analysis in language testing research: A handbook.* Cambridge, UK: Cambridge University Press.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219–244. doi:10.1177/0265532208101006.

Jarodzka, H., & Brand-Gruwel, S. (2017). Tracking the reading eye: Towards a model of real-world reading. *Journal of Computer Assisted Learning*, *33*(3), 193-201. doi:10.1111/jcal.12189.

Johnston, P. J. (1984). Assessment in reading. In P.D. Pearson, R. Barr, M. Kamil and P. Mosenthal (Eds.), *Handbook of reading research* (2nd ed.) (pp. 147–182). New York: Longman.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329-354. doi: 10.1037/0033-295X.87.4.329.

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge, UK: Cambridge University Press.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge: Cambridge University Press.

Kintsch, W. (2004). The Construction-Integration model of text comprehension and its implications for instruction. In R. Ruddell & N. Unrau (Eds.) *Theoretical models and processes of reading.* (5th ed.). Newark, DE: International Reading Association.

Koda, K. (2005). *Insights into Second Language Reading: A cross-linguistic approach*. Cambridge, UK: Cambridge University Press.

Lim, H. J. (2014). *Exploring the validity evidence of the TOEFL IBT reading test from a cognitive perspective* (Unpublished PhD Thesis). Michigan State University, Michigan, USA.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*(4), 207–218. doi: 10.1207/s15326985ep3404_2.

McCray, G., & Brunfaut, T. (2018). Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. *Language Testing*, 35(1), 51-73. doi:10.1177/0265532216677105.

McNamara, D. S. (2007). *Reading comprehension strategies: Theories, interventions, and technologies.* New York: Lawrence Erlbaum Associates.

McNamara, D., Ozuru, Y., Best, R., O'Reilly, T. (2007). 4-pronged comprehension strategy framework. In McNamara, D. S. (2007). *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 465-497). New York: Lawrence Erlbaum Associates.

McNeil, L. (2012). Extending the compensatory model of second language reading. *System: An International Journal of Educational Technology and Applied Linguistics, 40*(1), 64-76. doi:10.1016/j.system.2012.01.011.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012–1027. doi: 10.1037/0003-066X.35.11.1012.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), E*ducational measurement* (3rd ed.) (pp. 13- 103). New York: American Council on Education & Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi:10.1037/0003-066X.50.9.741.

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale,* 67(3), 215-227. doi:10.1037/a0032918.

Paulson, E. J., & Henry, J. (2002). Does the degrees of reading power assessment reflect the reading process? An eye-movement examination. *Journal of Adolescent & Adult Literacy*, *46*(3), 234-244.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading comprehension test performance. *Language Testing*, *20*(1), 26–56. doi: 10.1191/0265532203lt243oa.

Phakiti, A. (2008). Strategic competence as a fourth-order factor model: A structural equation modeling approach. *Language Assessment Quarterly*, *5*(1), 20. doi:10.1080/15434300701533596.

Pollatsek, A., Raney, G. E., LaGasse, L., and Rayner, K. (1993). The use of information below fixation in reading and in visual search. *Canadian Journal of Experimental Psychology*, *47*(2), 179–200. doi: 10.1037/h0078824.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading.* Hillsdale, NJ: Lawrence Erlbaum.

Prince, P. (2014). Listening comprehension: Processing demands and assessment issues. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.). M*easuring L2 proficiency: Perspectives from SLA* (pp. 93-108). Bristol: Multilingual matters.

Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modeling approach*. Cambridge: Cambridge University Press.

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.

Rayner, K. (1997). Understanding eye movements in reading. *Scientific Studies of Reading, 1*(4), 317-339. doi:10.1207/s1532799xssr0104_2.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422. doi: 10.1037/0033-2909.124.3.372.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology, 62*(8), 457–506. doi:10.1080/17470210902816461.

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading, 10*(3), 241–255. doi: 10.1207/s1532799xssr1003_3.

Rayner, K., and McConkie, G.W. (1976). What guides a reader's eye movements? *Vision Research, 16*(8), 829–837.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading.* Englewood Cliffs, NJ: Prentice Hall.

Rayner, K., Pollatsek, A., Ashby, J., & Clifton, C. (2012). *The psychology of reading* (2nd ed.). New York: Psychology Press.

Rayner, K., Reichle, E. D., & Pollatsek, A. (2005). Eye movement control in reading and the E-Z Reader model. In G. Underwood (Ed.), Cognitive processes in eye guidance (pp. 131 – 162). Oxford: Oxford University Press.

Rayner, K., & Sereno, S. C. (1994). Eye movements in reading: Psycholinguistic studies. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 57-82). San Diego, CA: Academic Press.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye- movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences, 26*(4), 445 – 526. doi: 10.1017/S0140525X03000104.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E-Z Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review, 16*(1), 1 – 21.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, *40*(2), 163-184. doi:10.1111/j.1745-3984.2003.tb01102.

Rupp, A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, *23*(4), 441–474. doi:10.1191/0265532206lt337oa.

Sarnaki, R. E. (1979). An examination of test-wiseness in the cognitive domain. *Review of Educational Research, 49*(2), 252-279. doi: 10.3102/00346543049002252.

Shohamy, E. G. (2001). *The power of tests: A critical perspective on the uses of language tests*. Harlow, England: Longman.

Smith, F. (1994). *Understanding reading: A psycholinguistic analysis of reading and learning to read (*5th ed.). Hillsdale, NJ: Erlbaum

Solheim, O. J., & Uppstad, P. H. (2011). Eye-tracking as a tool in process-oriented reading test validation. *International Electronic Journal of Elementary Education*, *4*(1), 153–168.

Spivey, M., Richardson, D., & Dale, R. (2009). The movement of eye and hand as a window into language and cognition. In E. Morsella & J. Bargah (Eds.), *Oxford handbook of human action* (pp. 225-248). New York: Oxford University Press.

Stanovich, K. E. (1984). The interactive-compensatory model of reading: A confluence of developmental, experimental, and educational psychology. *Remedial and Special Education*, *5*(3), 11–19. doi:10.1177/074193258400500306

Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers.* New York: Guilford Press.

Staub, A., & Rayner, K. (2007). Eye movements and on-line comprehension processes. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 327-342). Oxford: Oxford University Press.

Tian, S. (2000). *TOEFL reading comprehension: Strategies used by Taiwanese students with coaching -school training* (Unpublished PhD thesis). Teachers College, Columbia University, New York.

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

Ünaldı, A. 2004. Construct validation of the reading subskills of the Boğaziçi University English Proficiency Test (Unpublished PhD Thesis). Boğaziçi University, Istanbul, Turkey.

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

Weir, C, Hawkey, R, Green, T and Devi, S, (2009). The cognitive processes underlying the academic reading construct as measured by IELTS. *IELTS Research Reports Volume 9,* pp. 157–189. Retrieved from https://www.ielts.org/-/media/research reports/ielts_rr_volume09_report4.ashx.

Weir, C, Hawkey, R, Green, A, Unaldi, A and Devi, S, (2012). The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university. In L. Taylor and C. Weir (eds.), *IELTS Collected Papers 2: Research in reading and listening assessment. Studies in language testing.* (pp. 37–120). Cambridge: Cambridge University Press.

Wu, A. D., & Stone, J. E. (2016). Validation through understanding test-taking strategies: An illustration with the CELPIP-general reading pilot test using structural equation modelling. *Journal of Psychoeducational Assessment, 34*(4), 362-379. doi:10.1177/073428291560

Yang, P. (2000). *Effects of test-wiseness upon performance on the test of English as a foreign language* (Unpublished PhD thesis). University of Alberta, Edmonton, Alberta, Canada.