COMPARING L2 LEARNERS' STRATEGY USE
IN LITERAL VS. INFERENTIAL READING:
A COGNITIVE VALIDITY STUDY THROUGH EYE-TRACKING

DİDEM TUĞÇE ERDEM

BOĞAZİÇİ UNIVERSITY

2015

COMPARING L2 LEARNERS' STRATEGY USE
IN LITERAL VS. INFERENTIAL READING:
A COGNITIVE VALIDITY STUDY THROUGH EYE-TRACKING

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Master of Arts

in

English Language Education

by

Didem Tuğçe Erdem

Boğaziçi University

2015

Comparing L2 Learners’ Strategy Use

in Literal vs. Inferential Reading:

A Cognitive Validity Study through Eye-Tracking

The thesis of Didem Tuğçe Erdem

has been approved by:

Assoc. Prof. Gülcan Erçetin ______________________________
(Thesis Advisor)

Assist. Prof. Sibel Tatar ______________________________

Assist. Prof. Oya Özemir ______________________________
(External Member)

June 2015

DECLARATION OF ORIGINALITY

I, Didem Tuğçe Erdem, 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.…………………………………………...........

ABSTRACT

Comparing L2 Learners’ Strategy Use in Literal vs. Inferential Reading: A Cognitive Validity Study through Eye-Tracking

In language testing, a test's cognitive validity is assessed in terms of the match between cognitive processes elicited from a reader and the processes which a test taker would use in non-test conditions. A reading test should require readers to go through levels of comprehension for enhanced cognitive validity. Herein, the importance of careful examination of the cognitive processes of readers taking a reading test should be examined closely. As such, this study was motivated by the need to determine if eye movements could provide valuable information about test takers' cognitive processes as they answer literal and inferential questions in a standardized reading test. To this end, the participants’ eye movements in specified areas of interest were analyzed in detail using eye-tracking methodology considering its precedence over other techniques. Results of the study displayed that eye movements of competent and experienced readers do not differ showing they go through similar cognitive processes. Also, it was seen that there is a significant difference in the eye movements in certain interest areas but not in others while answering literal vs. inferential questions depending on going through levels of comprehension and using variety of reading strategies, substantiating the cognitive validity of the test. This study is important in terms of its exploratory findings and methodology as only a few studies in educational research examine cognitive processing and validity in second language through eye-tracking.

ÖZET

İkinci Dilde Okuma Sürecinde Soru Türlerinin Strateji Kullanımına Etkileri: Göz Hareketlerini İzleme Yöntemiyle Bulgulama

Dilde ölçme ve değerlendirmede bir testin geçerlilik ölçümünde doğal okuma ortamında gerçekleştirilen bilişsel davranışların test ortamında gerçekleştirilenlerle eşleşmesi temeldir. Bu sebeple, okuma becerilerini ölçmek için hazırlanmış bir dil testinin geçerliliğinin onanması için testi alan öğrencilerden farklı seviyelerdeki bilişsel davranışları sergilemeleri beklenmektedir ve dolayısıyla bu süreçlerin dikkatlice incelenmesi gerekliliği kendini göstermektedir. Bu da standart bir okuma testindeki metin odaklı ve çıkarsama odaklı soru türlerinin cevaplanması esnasında kaydedilen göz hareketlerini ve gerçekleştirilen bilişsel davranışları incelemeyi hedefleyen bir araştırmayı gerekli kılmıştır. Hedeflenen sonuca ulaşmak amacıyla katılımcıların farklı hedef ilgi alanlarındaki göz hareketleri ve bu hareketlerin örüntüleri, diğer tekniklere üstünlüğü düşünülerek göz hareketlerini izleme yöntemi ile detaylıca incelenmiştir. Araştırma sonuçları, yetkin okuyucuların gerçekleştirdikleri bilişsel süreçlerde farklılıklar olmadığını göstermiştir. Ayrıca, kullanılan testin bilişsel geçerliliğini doğrulayarak metin odaklı ve çıkarsama odaklı soruların cevaplanması sırasında göz hareketlerinin bazı hedef ilgi alanlarında istatistiksel olarak anlamlı bir farklılık gösterdiği, ancak bu farklılığın diğer alanlarda oluşmadığı gözlemlenmiştir. Göz hareketlerini izleme yöntemini kullanarak ikinci dilde okuma sürecinde gerçekleşen bilişsel süreçlerde bulgulama yapan ve testin bilişsel geçerliliğini sorgulayan eğitim araştırmalarının sayıca azlığı, bu araştırmanın metodolijisinin ve keşifsel sonuçlarının önemini vurgulamaktadır.

ACKNOWLEDGEMENTS

This master thesis has been a long, enlightening, and inspiring experience for me, and I would like to express my endless appreciation to those who helped me complete it. I would first like to express my thanks to the members of my thesis committee, Assoc. Prof. Gülcan Erçetin, Assist. Prof. Sibel Tatar, and Assist. Prof. Oya Özemir for their perceptive and illuminating comments and suggestions on this study.

My deepest thanks are to my thesis advisor Assoc. Prof. Gülcan Erçetin for her constructive guiding light throughout the writing process of this graduate study. Without her support and elucidative criticism on the research design and analyses, it would not have been possible to finish my thesis. I would also like to send my warmest thanks to Prof. Cem Alptekin for his academic support, and to Assist. Prof. Anthony John Pavlik for encouraging his students to take part in my study.

I am also indebted to Ümit Aslan for his technical assistance on the eye tracker. I owe special thanks to Çağla Nikbay and Hande Karakaplan for their cooperation on the graduate courses that we have taken together. I further extend my indebtedness to my friends for life, Elif Sempachedinoglou and Cafer Can Yılmaz, for their precious friendship and support since the beginning of undergraduate years.

Above all, I would like to thank my mother, Zeynep Erdem, father, Necdet Erdem, brother, Tuğberk Erdem, and fiancé, Burak Büyükgökçesu, for the unconditional love, emotional support, endless trust, and eternal motivation they have provided. I feel very lucky to be a part of this beloved family.

In conclusion, I would like to thank all of my instructors who have helped me to broaden my horizons, and instigated me throughout my educational life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AoI: Area of interest

CO: Correct option

EM: Eye movement

FPRT: First pass reading time

INCO: Incorrect options

MCQ: Multiple-choice question

NDRT: Reading subset of Nelson-Denny Reading Test

QS: Question stem

TFC: Total fixation count

TIR: Text irrelevant

TR: Text relevant

TRT: Total reading time

CHAPTER 1

INTRODUCTION

1.1 Background and purpose of the study

The validity of a test depends on the interpretation of the correct responses to questions, thus, the responses that are accepted as correct and the processes underlying these responses are the important factors in terms of test validity (Alderson, 2000). These factors necessitate the examination of the cognitive processing needed for completion of the task because, in language testing, a test is valid if the mental processes that a test elicits from a candidate resemble the processes that the test taker would use in non-test conditions (Field, 2013).

Reading comprehension is a complex process, and in non-test conditions, it requires readers to possess hierarchically ordered lower-level skills such as phonological awareness, alphabetic understanding, and fluency (Dole, Duffy, Roehler, & Pearson, 1991), and connect them with other higher-level skills because automaticity, background knowledge and schema construction, knowledge of text structures, and the capacity of different memory structures help general reading comprehension (Basaraba, Yovanoff, Alonzo, & Tindal, 2013). With respect to this, cognitive scientists agree that readers go through literal, inferential and evaluative comprehension levels, referred to as levels of comprehension, while reading in non-test conditions (Graesser, Millis, & Zwaan, 1997; Kintsch, 1988, 1998; Lorch & van den Broek, 1997). Thus, in test conditions, a reading test which is prepared to assess the academic language proficiency of readers is accepted as cognitively valid if it requires readers to go through these three levels of comprehension (Bax & Weir, 2012; Weir, 2005).

Also, proficient readers use various kinds of reading strategies (e.g., skimming, scanning, search reading, careful reading) in academic contexts (Khalifa & Weir, 2009; Urquhart & Weir, 1998), for which reason a reading test designed to assess readers' proficiency level in reading should also necessitate the use of these reading strategies in order to have cognitive validity (Bax & Weir, 2012).

Along these lines it is clear that to be able to determine whether a reading test is cognitively valid, requiring test takers to go through levels of comprehension and use various kinds of reading strategies, in-depth analysis of cognitive processing required for task completion is necessary.

Since most reading tasks are limited to the access of correct and incorrect responses to questions and much information about readers' thought processes is lost, resulting in a lack of evidence for the cognitive validity of a reading task (Dillon, 1997; Pellgrino, Chudowsky, & Glaser, 2001), different methods have been utilized by researchers to shed a light on cognitive processes of readers. To capture information about the thought processes of readers, techniques such as "stimulated recall", where readers are interviewed after a reading task and asked to recall their thoughts while they were completing the task or "protocol analysis", where readers are asked to "think-aloud" during task completion (Ericsson & Simon, 1993), have been commonly used as off-line techniques. However, these techniques have two significant weaknesses. First, they produce logistical problems as it is not feasible to interview each reader one-on-one for each reading task. Second, studies have shown that these techniques are at risk of expectancy effects, which occurs when a participant expects a given result or reports the expected result, and variability in outcome depending on the time the verbal protocol is collected as participants are at risk of forgetting to report some of the processes that they went through while reading (Hayes, White, Bissett, 1998; Kusela & Paul, 2000). For instance, retrospective protocols might not be complete, as the readers may not remember, or verbal protocols may be distracting (Tai, Loehr, & Brigham, 2006).

Because of these weaknesses, many researchers are interested in the immediate, on-line reading processes (Balota, Flores d'Arcais, & Rayner, 1990; Besner & Humphreys, 1991; Marslen-Wilson, 1989). To explore cognitive processes during silent reading, researchers have developed and used many techniques such as "word-by-word processing" (participants press a button to control the rate of presentation), "rapid serial visual presentation", called "RSVP", of sentences (subjects are exposed to words at a set rate in the same spatial location), and "completion responses" (subjects read a text silently and make a standard word recognition response, such as naming and lexical decision, to a subsequent target word). However, these techniques are not without their weaknesses as all of them are unnatural, and the reading rate emerging from these paradigms is often different from normal silent reading rate because of disruption of the flow of reading (Rayner & Sereno, 1994). Measuring the total amount of reading time for a larger segment of text (such as phrase, clause, or sentence), which is a variation of "word-by-word" reading paradigm, is another technique to examine cognitive processes in reading; however, although it provides natural reading, it is unable to register the exact time necessary to process individual words, leading to inaccurate indication of moment-to-moment processes (Rayner & Sereno, 1994).

Considering these weaknesses, researchers have focused more on EM measurement to explore cognitive processes in reading (Rayner & Sereno, 1994) as it does not disrupt normal reading rate and comprehension (Rayner & Sereno, 1994; Tinker, 1939), and it provides information on moment-to-moment processes in reading (Rayner & Sereno, 1994). In brief, words, sentences or long texts are displayed on a computer screen, and readers' eye movements (EMs) and fixations are recorded via eye-tracking while they read. An infrared beam is projected to the cornea of one eye, and a video camera reflects and records this light. Then, the corneal reflection of the light source is measured relative to the location of the pupil center (Duchowski, 2003). Although psycholinguists have used EM data gathered with the use of this tool, eye-tracking was not utilized by educational researchers until the present era as a moment-to-moment indication of cognitive processes (Duchowski, 2003; Rayner; 2009).

Educational researchers have used eye-tracking mainly to analyze word processing (Hyönä and Pollatsek, 1998; Inhoff & Rayner, 1986; O'Regan, 1979; Rayner & Duffy, 1986), inferences (O'Brien, Shank, Myers, & Rayner, 1988; Rayner & Sereno, 1994; Singer, 1994), syntactic processing (Frazier & Rayner, 1982; Rayner & Frazier, 1987), and global discourse processing (Blanchard & Iran-Nejad, 1987; Kaakinen, Hyönä, & Keenan, 2003) and these studies have been conducted mostly in a first language (L1). For that reason, there are only a handful of eye-movement studies in second languages (L2) investigating global discourse processing (O'Brien de Ramirez, 2008) and cognitive validity (Bax, 2013; Bax & Weir, 2012).

Therefore, this study attempts to investigate if test takers' cognitive processes as reflected through EMs differ between literal and inferential questions while taking a reading test in the L2, to gather evidence about the cognitive validity of the test.

Eye-tracking methodology was employed in the current study considering its advantages over other on-line methods. First, it might provide a means to observe the cognitive processes of the participants since EMs are moment-to-moment indicators of cognitive load, i.e. difficulties and easiness of processing which readers experience during reading (Hyönä & Lorch, 2004; Rayner & Sereno, 1994), and this provides the opportunity to analyze individual differences between readers at a very high level of detail (Bax & Weir, 2012). Second, eye-tracking methodology is nonintrusive, which means it does not disrupt the natural test taking or reading process with another task unlike other off-line or on-line methods (Rayner & Sereno, 1994; Tinker, 1939). Third, it lets readers be free to examine any part of the text in any order within a display screen, (Dussias, 2010; Hyönä & Lorch, 2004). In this respect, to have a detailed account of EM behaviors of the participants answering literal and inferential questions, three measures of EMs were used: the first pass reading time (FPRT), the total fixation count (TFC), and the total reading time (TRT). To complement these three measures, the sequence of the first 10 fixations was also investigated.

The following chapters are organized as follows: Chapter 2 will explain cognitive validity in reading. Chapter 3 will shed light on eye-tracking methodology, its use in reading studies in the L1 and L2, and the use of eye-tracking to study cognitive validity by reviewing the related research. Chapter 4 will give information about the methodology and design of the study. The results of the study will be displayed in Chapter 5, and these results will be discussed in detail in Chapter 6. Finally, conclusions drawn from the study will be reported in Chapter 7.

1.2 Definition of key terms

Eye movements (EMs): The term refers to eye movements or saccades indicating where the eyes move next as well as eye fixations indicating where and for how long the reader is looking at part of a text such as a word, a sentence, or a paragraph.

Nelson-Denny Reading Test (NDRT): The reading comprehension subset of a standardized test designed to measure reading ability of L1 and L2 readers.

Multiple-choice question (MCQ): The form of assessment in which responders are asked to select the best possible answer out of the choices from a given list.

Text relevant (TR): The part of the text that is relevant to a question.

Text irrelevant (TIR): The part of the text that is irrelevant to a question.

Question stem (QS): The stem of a question.

Correct option (CO): The correct option of a MCQ.

Incorrect options (INCO): The distractors of a MCQ.

Area of interest (AoI): It is the target area in a display screen and it will consist of a word, a sentence, a paragraph, or a whole text. AoI is adjusted considering a suitable margin of error to allow for individual variation in fixation. The TR, TIR, QS, CO, and INCO constitute the five AoIs in this study.

Fixation: The focus of one eye on a particular point in the text for at least 100 milliseconds.

First pass reading time (FPRT): The duration of the first forward fixation, which will be the only fixation or the first of several forward fixations, within the interest area.

Total fixation count (TFC): The frequency count of all individual fixations within a given AoI, showing how many times the target area was fixated.

Total reading time (TRT): All fixations observed within a given AoI, indicating how much time the participant spent reading the target in that area.

CHAPTER 2

TEST VALIDATION

Alderson (2000) states:

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

This quote emphasizes the importance of understanding the trait that is being measured, as it requires an insight into the cognitive processing needed for completion of the task.

For that reason, in language testing, mental processes that a test elicits from a candidate are considered to resemble the processes that the test taker would use in non-test conditions. Thus, “similarity of processing”, “comprehensiveness”, and “calibration” must be questioned while considering the validity of test items (Field, 2013) as these considerations, related to Messick’s (1989) notions of “construct-irrelevant variance” and “construct under-representation” from a cognitive processing perspective, entail the need for better understanding of test takers’ cognitive processing in language tests.

Considering the need to explore the potential contribution of eye-tracking to help assess the cognitive validity of reading test items, this chapter discusses issues related to levels of comprehension in reading, cognitive validity in academic reading, and eye-tracking studies conducted to study test validation and cognitive processes in reading.

2.1 Levels of comprehension in reading

Although reading comprehension is accepted as “the essence of reading” (Durkin, 1989), a recent research by the National Reading Panel has shown that researchers have paid attention to reading comprehension only in the last 30 years (National Institute of Child Health and Human Development [NICHD], 2000), probably because of its being more cognitively complex than the precursor skills required to facilitate and support it (Basaraba et al., 2013).

Reading comprehension, generally, can be defined as the ability to extract meaning or learn from a text (Rupley & Blair, 1983; Snow, 2002). Although this definition necessitates the acceptance of a simplistic view of reading comprehension which is a skill resulting from the independent, sequential development of hierarchically ordered lower-level skills such as phonological awareness, alphabetic understanding, and fluency (Dole et al., 1991), these basic level reading skills work in combination with other skills such as automaticity, higher-level language comprehension processes, background knowledge and schema construction, knowledge of text structures, and the capacity of different memory structures to aid general reading comprehension (Basaraba et al., 2013). Thus, text comprehension has been modeled by cognitive scientists as the construction of a multilevel mental representation by readers (Graesser et al., 1997; Kintsch, 1988, 1998; Lorch & van den Broek, 1997). In other words, the development of reading comprehension has been accepted as the result of emerging expertise with lower-level and higher-level reading skills considering the different complexity levels that necessitate readers to interact with a text to different degrees (Dole et al., 1991; Kintsch & Rawson, 2005). Within this perspective, readers should first be engaged in tasks of literal comprehension (Herber, 1970) which require readers to retrieve explicitly stated information from the text (Carnine, Silbert, Kame’enui, & Tarver, 2010) before being engaged in inferential comprehension tasks which require readers to understand relationships that might not be explicitly stated but crucial to understand the text (Applegate, Quinn, & Applegate, 2002) or evaluative comprehension tasks that require readers to analyze and critically interpret the text depending on prior knowledge (Basaraba et al., 2013). These tasks impose different cognitive demands on readers and require varying degrees of interaction with the text (Herber, 1970). They are also based on the type of information, i.e. textual or background information, the reader is expected to contribute to the types of questions in reading comprehension assessment (Leu & Kinzer, 1999; Rupley & Blair, 1983).

Readers should extract explicitly stated information in the text to be able to reach the first comprehension level, literal comprehension (Carnine et al., 2010), and this depends on readers’ word-level processing skills, or their ability to accurately recognize individual words and understand the meaning created by the combination of words into propositions and sentences (Perfetti, Landi, & Oakhill, 2005). Although word-level processing abilities are not adequate for global comprehension alone (National Research Council [NRC], 1998), these abilities are required to make deeper interactions with the text (Basaraba et al., 2013). Hence, test developers and instructors must realize that literal understanding is a building block for more advanced comprehension skills which must be examined to see the growth in readers’ performance (Kintsch & Rawson, 2005; Nation, 2005).

To reach the second comprehension level, inferential comprehension, readers interact with a text to make inferences about meaning that are not explicitly stated (Applegate et al., 2002). At this stage, readers are expected to manipulate the information in the text to understand the relationships among main ideas and details, and to use that information to draw conclusions about what the author actually wants to convey (Vacca et al., 2009), to complete omitted details, and/or to elaborate on the explicitly stated information (Dole et al., 1991). Thus, readers need to "read between the lines" while interacting with the text (Carnine et al., 2010), and this leads readers to construct a situation model of text (Graesser, Singer, & Trabasso, 1994; Kintsch & Rawson, 2005; Perfetti, 1999). This situation model requires readers to reach both literal and inferential understanding, and to apply their own background knowledge and prior experience to the text to aid or augment understanding (Basaraba et al., 2013).

Moving from literal to inferential level increases the cognitive load; thus, working memory has a key role in inferential comprehension. Moreover, language proficiency, reading skills, precise understanding of the requirements and goals of the reading task, and background knowledge pertinent to the text topic affect inferential comprehension (van den Broek, Lorch, Linderholm, & Gustafson, 2001). For example, a study by van den Broek et al. (2001) shows that younger readers might face more difficulty answering questions requiring inferential thinking because of their less automated basic reading skills. Also, Zwaan and Brown's (1996) study demonstrates that L2 readers engage in fewer higher-level processes as inference production compared to their performance in L1 reading, and also, they produce fewer associative and elaborative inferences to improve their text understanding while reading in the L2 compared to reading in the L1 due to lack of proficiency.

The last level, evaluative comprehension, requires readers to understand literal information, make interpretations about the author's intended meaning and/or understand the relationships between the details in the text to reach inferential comprehension, and analyze or evaluate given information in the text using prior knowledge or experiences (McCormick, 1992) or knowledge gathered from another source (Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001).

The Kintsch' Information Processing Theory (1998) also postulates three levels of mental representation in relation to the depth of processing. First, the exact wording of phrases (surface representation) must be kept temporarily active while the propositional representation of the text is constructed by the reader. Then, the reader gradually constructs a network of propositions that captures both the local and global coherence relations conveyed by the author by focusing on linguistic devices (e.g., anaphoric reference) that signal relationship types between concepts and propositions expressed in the text (Kintsch, 1998). Thus, reading has been characterized as consisting of microprocessing (local level processing) and macroprocessing (global level processing) which are two independent but complementary levels of processing (Kintsch, 1998; Kintsch & van Dijk, 1978; van Dijk, 1980). Locally, text comprehension necessitates readers to process the relations among the individual propositions conveyed in a text; however, globally, readers must detect the main ideas and points in a text and represent those macropropositions' relations to one another and to the subordinate propositions they dominate. These two levels of processing complement each other. Macroprocesses function on the microstructure representation to generate macropropositions that generalize and summarize the content of the microstructure, and the macrostructure representation is accessible to the microprocesses to help processing of local coherence relations (Kintsch & van Dijk, 1978; van Dijk, 1980). Additionally, with the support of background knowledge related to the topic of the text, readers might construct a representation of the situation described in the text (Kintsch, 1998).

At the literal level of reading comprehension, which is commonly defined as the reader's ability to "gain meaning directly from the print" (Walker, Munro, & Rickards, 1998, p. 88), readers decode surface code features, text-base meanings explicitly stated in the text and the connecting devices which bind these text components locally. This level of comprehension falls short to point to what authors mean although it can reflect what they say. For that reason, literal reading has been accepted as failing to give a deep understanding of text content (King, 2007), and has been associated with the performance of unskilled readers, who are considered as being incapable to go beyond the information included in a text (Walker et al., 1998).

The theory of levels of comprehension has also been supported by many studies (see Basaraba et al., 2013; Davey, 1988; McCormick, 1992; Snider, 1988). For Snider's (1988) study, junior high students with learning disabilities answered multiple choice questions (MCQs) related to 24 texts. Three types of questions were classified as textually explicit (literal), textually implicit (inferential), and scriptally implicit (evaluative). It was aimed to see if students' performance changed depending on question type. Results indicated that literal questions were the easiest to answer for the students, followed by inferential and then evaluative questions. Davey (1988) conducted a similar study with the same types of questions to see if the location of information, i.e. whether it is the whole text or a part of it, and inference type explained any variance in students' performance on a standardized reading comprehension measure. Results of the regression analyses showed that the location of response information explained 27% of the unique variance observed in struggling readers' performance and 12% of the variance in proficient readers. McCormick (1992) also observed that fifth grade students could answer, approximately, 70% of the literal questions correctly while they could answer only 61% of the inferential questions correctly, entailing that inferential questions were more difficult for them to answer. Moreover, the findings of the study (Basaraba et al., 2013) conducted to investigate the relative difficulty of items assessing literal, inferential and evaluative comprehension were consistent with prior research, that is literal items were easier than inferential ones and inferential items were less challenging than evaluative items for the participants taking the task.

2.2 Cognitive validity in academic reading

In real life reading, a wide range of cognitive processes is employed by a skilled and proficient reader in an academic context. Thus, an important issue in the development of language tests is whether the mental processes elicited by the test tasks resemble those employed in real life situations, i.e. whether they are cognitively valid (Weir, 2005). To be accepted as cognitively valid, the items of reading tests prepared to assess the academic language proficiency of students/readers should elicit the three levels of comprehension discussed above in order to emulate real-world academic reading processes. As Bax and Weir (2012) put it:

> . . . if a language test does not elicit from test takers the same type and level of cognitive processing as is used and expected in the real-world target situation, then it is not a valid instrument for assessing that area of linguistic behavior. (p. 3)

Regarding this, going through levels of comprehension is actualised by using different reading types and strategies as it is stated that reading types, i.e. expeditious vs. careful reading, indicate readers’ preferred ways of organizing and processing information (Messick, 1976). Additionally, McKay, Fischler and Dunn (2003) state that reading types applied while reading represent an individual’s typical or habitual mode of problem solving, thinking, perceiving, or remembering, and are accepted as stable characteristics of individuals. For that reason, Weir, Hawkey, Green and Devi

(2009) argue that reading models explaining readers' purposeful and strategic activities in an academic context and reading types specific to the academic context are important. The researchers state that:

> In general terms, the reading types covered [in an academic context] are expeditious reading, i.e. quick, selective and efficient reading to access desired information in a text (scanning, skimming and search reading), and careful reading, i.e. processing a text thoroughly with the intention to extract complete meaning from presented material. (p. 160)

Careful reading includes processing at sentential, intersentential, text and multi-text levels (Bax & Weir, 2012). Khalifa and Weir (2009) state that "careful local reading" necessitates processing at the decoding level until the basic meaning of a proposition is established. Some local inferencing might be required to build a mental model at the enriched sentence level. However, it does not demand integrating each new piece of local information into a larger meaning representation. For "careful global reading", the reader reads the text with a relatively high level of attention like studying a chapter or chapters in a core textbook at undergraduate level. The reader starts reading from the beginning of the text and continues through to the end, connects new information to a mental model, and then, creates a discourse level structure for the text which is appropriate for the reader's purpose. Also, the reader not only has to comprehend the micro- and macro-propositions but also how they are interconnected, which will require close and careful reading (even rereading of the whole text or at least the parts of it that are relevant to the purpose in hand). For that reason, tests or items focusing on sentence-level processing alone are not the best signs of academic reading ability (Bax & Weir, 2012), so their cognitive validity is questionable.

On the other hand, Rayner and Pollatsek (1989) argue that careful reading models are not enough to explain how readers handle expeditious reading behaviors such as skimming. Similarly, Carver (1992) and Khalifa and Weir (2009) mention that the speed and efficacy of reading is important as much as comprehension. Especially for reading for undergraduate studies, Weir et al. (2009) found in their study with a sample of university undergraduate students that quick, selective and efficient reading is more difficult than careful and efficient reading for a lot of readers.

Urquhart and Weir (1998) and Khalifa and Weir (2009) divide expeditious reading into three categories, as scanning, skimming and search reading. Scanning necessitates accurate decoding of a word or string of words at the local word level and it involves recognition and matching, so it is mainly used to achieve very specific reading goals such as looking for specific names, words/phrases, figures/percentages, dates of particular events or specific items in an index. Also, while scanning, a reader will not necessarily follow author's sequencing in a linear way. It is not even required to complete the reading of the sentence, to build the meaning beyond the clause, or to integrate the sentences into the structure of preceding text (Rosenshine, 1980).

Skimming, on the other hand, involves selective reading. That is, some sections of the text are either omitted or given very little attention; an effort is given to establish a macrostructure (the gist) on the basis of as few details from the text as possible (Urquhart & Weir, 1998). Skimming may help the reader decide quickly whether it is worthwhile to approach the text or parts of it again in a more careful fashion. According to Pugh (1978), skimming might:

> . . . also be used to obtain an overall impression of features of a text. For example, it may be used to glean surface information, to check on a writer's tone, or to discover how a writer structures a chapter. Related to discovering the structure is a further use of skimming, where the reader seeks ‘advance organization’ of what he is subsequently to learn in detail. (p. 54)

Thus, skimming is selective in terms of how much information readers decide to process: they may access words or possibly process entire sentences. The reader will pay his/her attention to the propositions that seem to be macro-propositional instead of redundancies by using their knowledge of text and genre which shows likely positions for macro-positions (e.g., first sentence of the paragraph). Nevertheless, due to the rapid nature of skimming, it is unlikely to reach a detailed meaning representation of the whole text. In order to arrive at a comprehensive and accurate text level structure, careful global reading is necessary (Urquhart & Weir, 1998).

Finally, search reading necessitates readers to sample the text, which can be words, topic sentences or important paragraphs, to find relevant information on a predetermined topic (Urquhart and Weir, 1998). Pugh (1978) states that in search reading:

> . . . the reader is attempting to locate information on a topic when he is not certain of the precise form in which the information will appear . . . the reader is not pursuing a simple visual matching task (as in scanning), but rather needs to remain alert to various words in a similar semantic field to the topic in which he is interested. It is true that the visual activity involved is similar to scanning in many ways. However, the periods of close attention to the text tend to be more frequent and of longer duration and, since information is more deeply embedded in the test, there is more observance of the way in which the author structures his subject matter and, hence, the linearity and sequencing. Information about the structure of the text may be used to assist in the search. (p. 53)

Supporting Pugh's definition, Guthrie and Kirsch (1987) put forward a model of search reading which is comprised of five components: “goal formation” – this necessitates reading the question and encoding its features to guide the search; “category selection” – the reader selects appropriate sections or subsections of the text for examination; “information extraction” – after extracting information, the reader needs to determine whether it fulfills his/her search goal; “integration” – the reader combines the extracted information with those parts previously searched or with his/her prior knowledge; “recycling” – the reader makes a judgment in terms of the adequacy of the extracted information in realizing the search goal. If it does not, the reader recycles through the previous processes until the search task is completed.

The process starts with a search for related vocabulary in the semantic field indicated by the task/item. Once the information required to answer a question has been quickly and selectively located, careful reading will take over and this may call for establishing propositional meaning at the sentence level, enriching propositions through inferencing, and thus requiring the reader to integrate information across sentences. If the information sought is extracted from within a single sentence, this might be best described as search reading local. On the other hand, where information from more than one sentence is required to produce an answer, then it is best viewed as search reading global (Urquhart and Weir, 1998), and search reading at the global level is the main expeditious reading skill for university students (Bax & Weir, 2012).

For cognitive validity, a reading task designed to assess readers’ proficiency level in reading should necessitate the use of all of these reading skills and types that are used in academic contexts. Khalifa and Weir’s (2009) processing model including integration of information, building a mental model of a text, as well as text-level comprehension, is a good start point to establish cognitive validity for reading comprehension tests/tasks, as this model accounts for the different reading types which readers use in academic life, the different processing levels which might be triggered, and the knowledge base essential to complete an assigned reading task effectively (Bax & Weir, 2012). Weir et al. (2009) state that the “subskills approach” to test reading is:

> . . . based on the assumption that it is possible to target particular types of item or test task to specific types of reading so that one item might target the ability

> to understand the meaning of an individual word in a text and another might target the ability to extract the overall meaning of a text within a very limited time frame (skimming). (pp. 162-163)

The researchers also emphasize:

> [T]he debate over subskills centered on the ability of expert judges to arrive at a consensus about what was being tested and the essential role of the candidate was largely overlooked. The majority of the studies paid surprisingly little attention to the cognitive processing required for candidates to carry out test tasks. (p. 63)

However, Alderson (2000) states that:

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

Thus, Bax and Weir (2012) point out that understanding of reading proficiency necessitates a deep analysis of cognitive processing required for task completion, and eye-tracking can be an effective tool to analyze cognitive processing and assess cognitive validity of a given task.

In conclusion, it is agreed by authorities that validity is the building block of an effective test, thus mental processes that a test elicits from test takers should be similar to those occurring in non-test conditions (Alderson, 2000; Bax & Weir, 2012; Field, 2013; Weir, 2005). To design a reading test which is cognitively valid in an academic environment, it is important to ensure that the test includes items assessing literal, inferential and evaluative comprehension which necessitate the use of all reading skills and types used in academic contexts (Bax & Weir, 2012; Dole et al., 1991; Khalifa & Weir, 2009; Kintsch & Rawson, 2005; Leu & Kinzer, 1999; Rupley & Blair, 1983; Weir et al., 2009). As a result, studies investigating levels of comprehension (Basaraba et al., 2013; Davey, 1988; McCormick, 1992; Snider, 1988; van den Broek et al., 2001; Zwaan & Brown, 1996) found similar results in that literal questions are easiest to answer while evaluative questions are the most difficult to answer by being cognitively more demanding.

CHAPTER 3

EYE MOVEMENTS IN READING

3.1 History of eye-tracking

The eye-tracking technology has been utilized for 135 years, and this period was divided into three by Rayner (1998).

The first period started in 1879 with Javal’s initial observations related to the role of EMs in reading (Huey, 1908), and ended in 1920. During this period, the basic characteristics of EMs themselves (i.e., saccades, EMs made from one fixation point to another) such as saccadic suppression (the suppression of vision during a saccade), saccade latency (the time necessary to set off EM), and the size of perceptual span (the region of effective vision) were the focus points (Rayner, 1978, 1998).

During the second period improving reading by training EM behaviors were undertaken with the effect of behaviorism (Rayner, 1998). Despite the classic work by Tinker (1946) on reading and by Buswell (1935) on scene perception, during this period the focus was on EMs per se without considering its relation to cognition. Little research was done with EMs to infer cognitive processing (Rayner, 1998).

The third era, which started in the mid-1970s, was characterized by technological advancements and by the assumption that “EM data reflect moment-to-moment cognitive processes.” (Rayner, 1998, p. 372). As Rayner (1998) puts it:

> A crucial point that has emerged recently is that eye movement measures can be used to infer moment-to-moment cognitive processes in reading . . . and that the variability in the measures reflects on-line processing. For example, there is now abundant evidence that the frequency of a fixated word influences how long readers look at the word. (p. 376)

Additionally, according to the eye-mind hypothesis proposed in that period, EMs are used as a measurement of ongoing mental process in reading. The hypothesis suggests that the focus of attention will be revealed by human gaze behavior when the visual environment is relevant to the task at hand (Just & Carpenter, 1980). For that reason, the hypothesis posits that human gaze behavior indicates the temporal change of visual attention and essential facets of information decoding and integration (Duchowski, 2003). Concordantly, Spivey, Richardson, and Dale (2009) state that EMs can be taken into consideration as good indicators of cognitive processes, and they define these EMs as windows into language and cognition. Salvucci and Goldberg (2000) also use the same metaphor, as they accept eye-tracking as "a window of observers' visual and cognitive processes" (p. 71). De Greef, Botzer, and van Maanen (2010) even take this into further by arguing that EMs are tools to read the mind, which is in parallel with the eye-mind hypothesis.

As eye-tracking technology provides supplementary opportunities to have insights into readers' actual, rather than reported, behaviors, with the developments in eye-tracking technology, it is possible to spot what readers are looking at moment-to-moment (Bax & Weir, 2012; Rayner, 2009) making available a detailed analysis of individual differences between readers at a very high level of detail (Bax & Weir, 2012).

3.2 Attributes of EMs in reading

Since the beginning of the use of eye-tracking in reading, EM data have provided insightful information on various aspects of reading process such as word processing, syntactic processing, and global discourse processing because eye-tracking measures draw on fixation sequence during which visual information intake is completed while new visual locations are brought to fovea, the center of the eye (Rayner, 1998).

Research has shown that readers see nothing when their eyes move throughout a text, and it is only possible to process textual or inferential information via fixations. During these fixations, the eyes can project foveal, parafoveal and peripheral regions to view information. As being the central area, the foveal region includes six to eight letters while the parafoveal region extends to about 15 to 20 letters. The third region, namely peripheral region, comprises everything in the visual field beyond the parafoveal region though it becomes more difficult to identify the words presented to locations away from the fovea, as the region concerned with processing details is the foveal region (Rayner & Sereno, 1994).

During reading, the eyes move in a series of jumps named “saccades” instead of moving smoothly across the text. However, they stay rather stable between these jumps, and these stable periods are termed “fixations”[1], when the eye dwells momentarily on a particular point. The number and duration of fixations are influenced by many textual and typographical variables “ . . . as text becomes conceptually more difficult, fixation duration increases, saccade length decreases, and the frequency of regressions increases.” (Rayner, 1998, p. 376), and task relevance (i.e., task relevant stimulus) has a boosting effect on the duration of individual fixations (Kaakinen, Hyönä, & Keenan, 2002). For instance, readers skip more words if they are speed-reading or skimming (Taylor, 1962). However, it is commonly accepted that the average saccade duration is 20-40 ms, and the typical fixation duration is 200-250 ms (100 ms minimum, 500 ms maximum) (Staub & Rayner, 2007). Some saccades might extend more than 20 characters; however, in

[1] Rayner (1998) classifies nystagmus, drifts and microsaccades as three types of EMs. Eyes are not precisely still while fixating, as very small tremors, called nystagmus, occur. In reading research, nystagmuses are considered as noise and are not included in scoring.

normal reading in English the eyes move about seven to nine letter spaces on the average saccade (Rayner, 1997; Staub & Rayner, 2007). Processing of new information in reading occurs only during fixations (Rayner, 2009; Staub & Rayner, 2007), as eyes are moving very fast across the stable visual stimulus that only a blur would be perceived during saccades (Rayner, 2009) though cognitive processing continues in most situations (Irwin, 1998; Irwin & Carlson-Radvansky, 1996). In summary, the eye fixation data show both endogenous and exogenous attention shifts, respectively (Amadieu, van Gogh, Paas, Tricot, & Mariné, 2009).

Though most of the saccades are forward (from left-to-right) while reading in English, about 10 to 15% of the saccades are regressions, which means 10-15% of fixations are backward (from right-to-left) through the line or movements back to already read lines (Rayner, 1978, 1998). Regressions will be short, suggesting processing difficulties specific to a word; however, longer saccades made to previous sentences in long texts imply processing difficulties and comprehension failures due to general text difficulty (Roberts & Siyanova-Chanturia, 2013).

It is also important to mention that skilled readers can progress approximately 240 words per minute (Just & Carpenter, 1987), and readers make predictions and inferences based on textual and contextual cues (Goodman, 1996), which makes the reading more efficient and better comprehension possible. And because skilled and unskilled readers' EMs (fixation duration, saccade length, frequency of regressions) are different, they could potentially be useful to compare better and worse readers (Bax & Weir, 2012). For example, more efficient readers are likely to have fewer but longer fixations, with longer saccades between them (Rayner, Pollatsek, Ashby, & Clifton, 2012), and they tend to be more "strategic" (Pang, 2008), so they are more likely to use longer saccades while locating target areas of a text.

Certain areas of texts are spotted and they can be experimentally manipulated in EM studies. These identified regions may contain a couple of words or sentences (Clifton, Staub, & Rayner, 2006). During reading, eyes fixate on some parts of a text for a certain time. These parts are considered as areas of the text being attended or AoI. Hence, fixation times and fixation durations on target areas can be calculated to understand cognitive processing (Rayner & Sereno, 1994). All in all, it can be argued that eye fixation, along with its key indicators such as the number of fixations, percentage of total fixations, and total fixation duration in the AoI, is an appropriate measure to study the time period necessary to acquire new information (Rayner, 2009). However, multiple fixations, regressions on certain areas or words are accepted as the proof of comprehension difficulties (Rayner & Pollatsek, 1989), because poor readers having comprehension difficulties show shorter saccades, longer fixations, and more regressions (Goltz, 1975; Griffin, Walton, & Ives, 1974; Heiman & Ross, 1974; Rubino & Minden, 1973). For that reason, when the target is larger than a word, the first-pass (the initial reading) and second-pass (rereading) reading time of the target area is frequently measured in reading research (Rayner, 1998).

In conclusion, eye-tracking has highly been preferred as a methodology to study the cognitive processing (Hyönä, 2010) as eye trackers have become more participant-friendly with their unobtrusive nature, and researcher-friendly with their ready-made analysis software packages that help draw sensible conclusions from rich datasets. Additionally, it is important to complement on-line eye-tracking data with off-line measures (e.g., metacognitive awareness of reading strategies inventory) to exploit the end product of learning (Winke, Gass, & Sydorenko, 2013).

3.3 Word processing

Frequency, word familiarity, morphology, lexical ambiguity and context are the factors that influence word processing, thus EM research focusing on word processing gather data about them.

Research has shown that, controlled for word length, frequency of a word affected the first fixation duration and gaze duration on the word (Inhoff & Rayner, 1986; Rayner & Duffy, 1986) with high-frequency words being skipped more frequently than low-frequency words (O’Regan, 1979; Rayner, Sereno, & Raney, 1996). Word familiarity is another factor affecting word processing in that the eyes spend less time on familiar words leading to shorter fixation durations (Chaffin, Morris, & Seely, 2001; Juhasz & Rayner, 2003; Williams & Morris, 2004).

The effect of morphology on word processing was investigated as well. Considering the argument that compound words are decomposed into their constituent morphemes as they are analyzed, Hyönä and Pollatsek (1998) manipulated the frequency of the first constituent while keeping the frequency of the whole word of transparent Finnish compound words constant, and they saw that fixation durations were affected by the frequency of the first constituent. Also, the same result was reached after studies on English compound words (Andrews, Miller, & Rayner, 2004; Juhasz, Starr, Inhoff, & Placke, 2003).

Lexical ambiguity is another factor affecting word processing. Specifically, fixation duration has been shown to be longer with words which have two meanings compared to unambiguous control words (Duffy, Morris, & Rayner, 1988; Rayner & Duffy, 1986; Rayner & Frazier, 1989; Sereno, O’Donnell, & Rayner, 2006).

Contextual factors have been manipulated and the effect of context on word processing has been investigated by researchers. Various studies show that when a word could be predicted from or constrained by the preceding context, fixation durations on the word decreased (Balota, Pollatsek, & Rayner, 1985; Zola, 1984). Furthermore, it is seen that high predictable words are skipped more frequently than low predictable words (Rayner & Well, 1996).

3.4 Inferences

Readers might infer information that has not been explicitly stated up to a given point in a text, and this is called elaborative inference (Rayner & Sereno, 1994). Models of discourse comprehension argue that relevant world knowledge is activated by readers to comprehend texts (see Glenberg, Kruley, & Langston, 1994; Kintsch, 1994; Singer, 1994). Thus, readers should combine their inferences (implicit knowledge) with what is actually stated in the text (explicit knowledge). As inferences have a critical role in comprehension and processing, it is significant to illustrate their existence empirically, which is possible with EM studies (Rayner & Sereno, 1994).

O'Brien et al. (1988) examined fixation time on a target word (e.g., knife) in the final sentence of a passage (1) on-line to explore processing while doing elaborative inferences.

(1) *He threw the knife into the bushes, took her money, and ran away.*

In the first condition, the target word "knife" was previously explicitly mentioned in the text (e.g., by adding the phrase *stabbed her with his knife*), and in the second condition the target word was previously strongly suggested (e.g., by adding the phrase *stabbed her with his weapon* to the passage), which necessitated inferencing to connect the word "weapon" to the target word "knife" occurring in the final sentence of the passage (1). No difference was observed between the conditions in terms of gaze duration on the target word, showing that the word “knife” was inferred from the preceding context by the readers. However, when the word “knife” was not strongly suggested by the text (e.g., by the phrase *assaulted her with his weapon*), gaze duration on “knife” which was in the final sentence was longer compared to conditions when the target word was explicitly mentioned or strongly suggested previously in the text. The researchers concluded that the longer gaze duration on “knife” was a result of memory search and cognitive load of the sentence.

Additionally, in order to find the effect of inference on processing, anaphora resolution, accepted as a type of bridging inference (Singer, 1994), was examined while gathering EM data by researchers. Researchers varied the distance between a pronoun and its antecedent by placing the antecedent near or far.

Rayner, Chace, Slattery and Ashby’s (2006) experiment was conducted to see whether EMs can inform the understanding of anaphor processing during silent reading of long passages of text. The experiment presented six sets of 36 paragraphs that were 150 words long and which were the same except for the lines containing the antecedent. The researchers focused on the embedded sentences that contained a target anaphor; half of the passages contained anaphors that were consistent with their antecedents (2a), and the other half contained inconsistent anaphors (2b).

(2a) *Alison decided to order some carrot sticks to snack on. The waiter brought her some water and the carrot sticks after only a few minutes.*

(2b) *Alison decided to order some celery sticks to snack on. The waiter brought her some water and the carrot sticks after only a few minutes.*

Also, the antecedent and the anaphor were near to each other (average 10-15 words intervening), at an intermediate distance (average 50-55 words intervening), or at a far distance (average 120-125 words intervening). The experiment had a 2 (anaphor:

consistent vs. inconsistent) x 3 (distance: close, middle, far) repeated measures design. 18 adult skilled readers of English read six paragraphs presented in random order in each of the six conditions at a normal pace, and answered yes/no questions that followed the paragraphs. At the end, the researchers found that inconsistencies in the near condition led to longer fixations on the inconsistent anaphor and more regressions from the words immediately following the anaphor back to the antecedent. Moreover, it was seen that increasing the distance between the anaphor and its antecedent reduced the probability of inconsistency detection. However, the researchers reported that the regression data were in conjunction with the findings that second pass reading time of the antecedent in the consistent condition, which was 32 milliseconds (msec.), was shorter than that in the inconsistent condition (66 msec.), F (1.17) = 10.04, p < .01, showing that even though readers might not be much accurate representing the antecedent across larger distances, they found and reread the antecedent, regardless of its distance from the anaphor.

The findings of the study conducted by Rayner et al. (2006) were in line with previous studies (see Ehrlich & Rayner, 1983; Garrod, Freudenthal, & Boyle, 1994; O'Brien, Raney, Albrecht, & Rayner, 1997) in which the distance between a pronoun and its antecedent was varied by placing the antecedent near or far. These studies showed that fixation durations were longer in the far condition and shorter in the near condition. For that reason, it was concluded that when the distance between the anaphor and its antecedent became further, fixation duration on the pronoun increased with the effect of cognitive processing difficulties.

3.5 Syntactic parsing

Eye-tracking research on syntactic parsing can be categorized into two groups. In the first category, temporarily ambiguous sentences between two syntactic structures are read by participants and gaze durations or regressive movements in the ambiguous area compared to control region. Research in the second category, however, focuses on semantic/pragmatic or syntactic anomalies in order to see the time and place of disruption effects.

In the first group, some studies tested the validity of “garden path” theory in syntactic parsing to determine readers’ structural analysis while dealing with ambiguity. Two principles, namely “minimal attachment” (Frazier & Rayner, 1982) and “late closure” (Rayner & Frazier, 1987), were tested to clarify participants’ parsing process of temporarily ambiguous sentences. Findings indicated longer fixations and more regressions in the ambiguous area due to cognitive load caused by processing demands. Also, other studies tested the effect of animacy and inanimacy on grammatical constraints in sentence processing, and it was found out that neither animacy nor inanimacy reduced the difficulty of ambiguous sentence processing as suggested by increased reading time on the problematic area (Clifton et al., 2003).

The second group of studies examined semantic and/or syntactic plausibility in the L1 (Braze, Shankweiler, Ni, & Palumbo, 2002; Deutsch & Bentin, 2001; Murray & Rowan, 1998; Ni, Fodor, Crain, & Shankweiler, 1998; Pearlmutter, Garnsey, & Bock, 1999; Rayner, Warren, Juhasz, & Liversedge, 2004). Although the results were mixed, meaning there was not a shared finding related to the type of EM patterns (e.g., first-pass effect, total time effect, regression) on the critical region, these studies showed that syntactic or semantic anomaly caused more fixation durations and/or regressive EMs.

3.6 Global discourse processing

There are not many EM studies that examine the processing of long texts. In Hyönä, Lorch, and Kaakinen's (2002) study, Finnish readers read two expository texts in Finnish at an approximate length of 1200 words. The organization of the texts was the same, both texts began with a brief introduction which was followed by the discussion of 10 or 12 different topics. These topics were categorized into two main sections of the same issue, and a heading was used to signal topic shift for each topic. Each topic consisted of two paragraphs, and they started with a topic sentence followed by supporting sentences. Depending on the readers' global processing, they were categorized as "fast linear readers", "slow linear readers", "topic structure processors" or "nonselective reviewers". The first group, "fast linear readers", did not make regressive EMs while the second one, "slow linear readers", made many forward fixations and reinspection of each sentence before reading the following sentence. Furthermore, "topic structure processors" mainly focused on topic headings, and "nonselective reviewers" made many regressive fixations on previous sentences. Consequently, the research illustrated the variety of readers' reading behaviors depending on their cognitive processing.

To determine whether a global processing would lead to occurrence of different EM patterns compared to the EM patterns observed during local processing, Blanchard and Iran-Nejad (1987) examined the EM patterns of skilled adult readers when they were reading stories with surprise endings. Participants' comprehension processes were manipulated at the most global level. Three stories, whose length ranged between 879 and 1041 words, were used in the experiment. There were experimental and control versions of each text. Experimental versions were the original surprise-ending stories. The control versions were identical to the experimental versions, except for the additional sentences added to the beginning of the stories to give away the surprising information. Before the sessions, the participants were told that they were required to summarize what they read, so they read the stories with a purpose. The participants read either a control or experimental version of each of the three stories, and each person received at least one of the three stories in the surprise-ending version. The lines of the texts were displayed on-line at a time without a set time limit; however, the participants could not go back to the line that they had already read and passed. The researchers observed that the mean reading times for selected lines of the control texts (2613 msec.) were shorter than the mean reading times for selected lines of the experimental texts (3401 msec.). Also, the fixation durations in the selected lines of the control texts (first pass: 202 msec., second pass: 200 msec.) were shorter than the fixation durations in the selected lines of the experimental texts (first pass: 212 msec., second pass: 212 msec.). Furthermore, it was found that the participants made 384 fixations during the first pass and 90 fixations during the second pass in the selected lines of the control texts while making 508 fixations during the first pass and 277 fixations during the second pass in the selected lines of the experimental texts. Last but not least, it was seen that the percentages of the second pass fixations in the selected lines of control texts 1, 2, and 3 were 10, 22, and 20 respectively, but they were 25, 42, and 33 in the selected lines of the experimental texts 1, 2, and 3, respectively. Thus, these results showed increased reading time for the text which introduced the surprising information, increased fixation durations and increased number of fixations in the surprise-ending part which was a result of the rereading of the surprise-ending section, and an increase in time spent on rereading the text. In the light of these results, the researchers concluded that even the most global levels of processing influence EM patterns.

To have more information on global discourse processing, prior knowledge effect on on-line processing of perspective-relevant and perspective-irrelevant information in expository texts was examined (Kaakinen et al., 2003). Finnish university students read two texts in L1 Finnish. The first text was about familiar diseases which the participants had much prior knowledge about, while the second text was about rare diseases about which they had little or no prior knowledge. Also, a reading perspective (e.g., the participants pictured themselves as elementary school teachers who needed to give pupils information on a specific disease) was set by the researchers to make pupils read the texts more carefully by paying attention to details and specific information. Throughout reading, participants' EMs were recorded, and then a recall task and a reading span task (RST) were administered. Findings indicated that compared to the familiar text, less information was recalled from the unfamiliar text, and also, longer fixation durations were measured for the unfamiliar text and for the sentences that are relevant to the participants' reading perspective. Moreover, according to Kaakinen et al. (2003) categorization, the low-span group read the sentences that were related to their reading perspective slowly by demonstrating longer first-pass fixation durations. However, only for the unfamiliar texts, longer first-pass fixation points were recorded for the high-span group illustrating that having a rich prior knowledge led them not to slow down their initial reading of relevant sentences. Furthermore, it was observed that the low-span readers made more regressive fixations on perspective-relevant than on perspective-irrelevant information showing their restoration of the relevant information to working memory; however, no significant perspective effect was found for the high-span readers. In line with the findings, it was concluded that reading span had an effect on the ability to use prior knowledge in reading process, and the EM patterns gave information about the participants’ cognitive processes, meaning longer fixation durations (e.g., first-pass fixation, total fixation, or regression durations) signaled cognitive processing difficulty.

Kaakinen et al.'s (2003) study was replicated by Burton and Daneman (2007) with the English version of Finnish texts. The study investigated the effect of epistemic knowledge instead of prior knowledge. It was found that having mature or naive epistemic knowledge did not have an effect on first-pass progressive fixations and first-pass rereading fixations. Nevertheless, it was also seen that, compared to metacognitively naive low-span readers, metacognitively mature low-span readers made more regressions; in other words, they were more involved in strategic backtracking. The occurrence of regressions as a compensatory strategy on more difficult target sentences in texts about unfamiliar diseases pointed out the cognitively loaded nature of these parts.

Rayner et al.’s (2006) experiment demonstrated that global discourse difficulty increases the duration of fixations, the number of fixations and the probability of fixations during silent reading of long passages of text. During the experiment, eye data for several measures (average fixation duration, TFC, and total fixation time for the passage) were collected from 16 native English speakers while they were reading long passages (32 passages of text that had a mean length of 564 words) of more or less difficult text. The text difficulty was determined by the ratings of 32 readers on a scale that ranged from 1 to 10, with 1 being the easiest and 10 being the most difficult, and with respect to ratings, passage difficulty ranged from 2.8 (relatively easy text) to 6.6 (moderately difficult text). Participants read the texts silently at their own pace to answer the four-choice comprehension questions following every passage. When the participants read the text, they pressed a button and the passage on the screen was replaced by a question. At the end, it was observed that the difficulty rating was significantly positively correlated with average fixation duration (which was 3 msec. longer for the difficult passages than for the easy ones), the TFC (25 more fixations on the difficult passages), and the total fixation time (8.1 sec. longer to read the difficult passages). Even though difficulty was negatively correlated with accuracy which was 6.6% better on the easy items, denoting poorer comprehension for more difficult passages, this correlation was not statistically significant, t (15) = 1.58, p > .05. The researchers concluded that fixation durations and the TFC are influenced by passage difficulty, and they also show that EM measures are sensitive to global passage difficulty.

In conclusion, it can be said that EM research results have given more insightful information than other methods used in reading research, and have provided evidence to support hypotheses on reading processes such as text comprehension models and individual differences in comprehension such as working memory, reading ability, and using metacognitive reading strategies.

3.7 EM research in L2 global discourse processing

EM studies investigating global discourse processing in the L2 is exceedingly limited. O'Brien de Ramirez (2008) compared global text processing in the L1 with the processing in the L2. Highly proficient late adult bilinguals read short stories in L1 French or L1 English and in L2 French or L2 English. The length of the stories varied from 1000 to 1300 words and consisted of 102 lines. Findings indicated that full text reading in English was 2.68 msec. faster than in French as an effect of word length difference between these two languages. Additionally, although it was found that reading speed in L2 English was slower than reading in L1 English in general, reading in L2 French was faster than reading in L1 French. Nevertheless, there was not a positive correlation between reading speed and fixation length, which means even though L2 readers of French read the story faster, the mean of their fixation duration was 17% longer than the mean fixation duration of L1 readers of French. Similarly, mean fixation durations of L2 English readers were 1.8% longer than mean fixation durations of L1 readers of English showing the fact that readers of L2 fixated the text more frequently, so it took more time to read the text. As a result, and more importantly, it can be said that longer fixation durations or more frequent fixations were observed in bilinguals' EMs while they were processing a text in L2.

3.8 Eye-tracking to unravel strategy use

The belief that human gaze behavior will unravel the focus of attention when the visual environment and the task at hand are related to each other, which supports the eye-mind hypothesis, led researchers to design eye-tracking studies in order to explore the strategies used during the process of reading (Just & Carpenter, 1980).

To understand the relationship between EMs and strategy use, Carpenter and Just (1989) recorded EMs of six low-span and six high-span participants while they completed a RST. They saw that the low-span readers read slower and scored lower on the task while the high-span readers read faster (although they spent more time on the sentence-final words compared to other group) and scored higher on the task. The researchers concluded that the better readers read faster and used their extra time to fulfill the requirements of the given task. Thus, the low-span readers moved on reading for meaning; however, the high-span readers processed sentences superficially to ease remembering sentence-final words as a strategy.

Kaakinen and Hyönä (2007) investigated strategy use in reading via gathering EM data in order to have an elaborated view of the time-course of processing of the test materials and they asked the participants' memory encoding strategies for the same task. Experimenter-paced Finnish version of Daneman and Carpenter's (1980) RST was used. Gaze duration and total fixation times for the first words of each sentence, middle-of-sentence words, and end-of-sentence "to-be-remembered" words were calculated for medium and low-span readers. In contrast to Carpenter and Just's (1989) results, they found out that there was no difference between the groups caused by the use of different on-line processing strategies as both groups processed the irrelevant parts (beginning and middle regions) in less time and the "to-be-remembered" word, which necessitates more cognitive load, in more time. To find more about strategy use, Kaakinen and Hyönä (2007) interviewed participants taking the same test in another study. They divided the responses into three categories as "semantic elaboration", which is producing sentences using the sentence-final words, forming relations between the sentence-final words and personal experience, or imagery, "silent rehearsal of the final words while reading the sentences", and "selective processing of the sentence", which is concentrating on the last word rather than paying attention to the meaning of the sentence. It was seen that the low-span participants mostly used rehearsal strategies while the high-span participants used semantic elaboration more, showing the fact that a reader needs to have cognitive resources to use effective strategies.

Van der Schoot, Vasbinder, Horsley, and van Lieshout (2008) examined two reading strategies used by 10-12-year old children while reading in the L1, and these strategies were differentiating important words from unimportant words and resolving anaphoric references. Readers' eye fixations on specific target words were recorded in order to study reading strategy use. At the end, it was seen that more successful comprehenders spent more time to process important words than to process unimportant words. However, it was also observed that it took them less time to find the antecedent of an anaphor. The findings of the study provided evidence for macrostructure building theory of construction-integration in reading comprehension (Kintsch, 1998) since according to Kintsch's (1998) theory readers should focus on macropropositions instead of concentrating on unimportant information to build a coherent mental model.

3.9 Eye-tracking to study test validation and cognitive processes in reading

Although eye-tracking has been used extensively by psycholinguists since the first era of EM research, it was not utilized by educational researchers until the present era as a moment-to-moment indication of cognitive processes (Duchowski, 2003; Rayner; 2009). Hyönä and Lorch (2004) state that as a method to study global text processing strategies, eye-tracking is effective since it allows investigation of on-line processing. It can collect different indices of processing simultaneously which is not possible to do via other on-line methods. Furthermore, eye-tracking allows readers to be free to examine any part of the text in any order within a display screen, and the normal reading is not disrupted as it is never interrupted with another task which is not possible in many on-line methods (e.g., probe procedures) (Dussias, 2010; Hyönä & Lorch, 2004).

Solheim and Uppstad (2011) state that eye-tracking provides its users with an alternative window into comprehension process and problem solving behavior and can facilitate a validation process. The relations of comprehension scores with actual behavior can be searched by the use of eye-tracking as it externalizes parts of the reading processes in the form of records of EM patterns. The researchers (2011) argue:

> 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. (p. 155)

For that reason, by looking at the percentage of time that is spent on reading relevant information, for example, or that is spent on the use of that information to answer the question, researchers will have an idea about the cognitive processes during reading to be used while assessing cognitive validity of tasks (Solheim & Uppstad, 2011).

In order to assess the cognitive validity of a reading test (computer-based Certificate in Advanced English (CAE) Reading Test), Bax and Weir (2012) investigated the cognitive processes utilized by L2 readers via eye-tracking. They aimed to see the extent to which test items elicited the range and level of cognitive processes expected of an advanced reading test designed to reflect real-world academic reading processes. Multinational students studying at Foundation level, Year 1 and Year 2 undergraduate levels participated in the study. First, the students took an edited version of the original CAE test. Two parts of the original test was omitted by the researchers in order not to face any difficulties in comparing the participants' behavior through eye-tracking and to be able to design a test which does not take longer than 30 minutes to be completed. This edited version of the original CAE test was divided into two parts. Part 1 included three themed short texts with two MCQs on each; and Part 2 consisted of one long text with a side scrollbar and related seven MCQs on it. The participants were given 30 minutes to answer 13 questions and finish the test. After reading each question, the participants answered seven onscreen retrospective questions to elicit immediate recall of the cognitive processes that were used in order to find whether readers had read globally or locally, carefully or expeditiously, or had used word-search strategies. The onscreen retrospective questionnaire aimed to obtain information about how the participants had approached the text and the questions, the particular cognitive strategies that were used when they tried to find the answer to the question, and whether they used local or global processing while answering the questions. Throughout the whole process, 35 participants' EMs were recorded. However, eye-tracking activities of six participants for chosen four questions and eye-tracking activities of four participants for chosen one question on CAE Reading test were analyzed. Total fixation duration, the FPRT and the TFC were calculated with the use of video, gaze plot and heat map data and automated statistical analysis. Regarding EM data, it was found that all of the participants had read each question carefully and they all read the question before reading the text (with the exception of one participant). Also, 92.9% of the participants used appropriate expeditious strategies to find the correct part of the text for each answer, and 96.4% of them had read all the options on all test items carefully (three fixations per option). Moreover, all participants except one spent longer time on the questions (average 30.57 seconds) than on the text (average 16.57 seconds), and they all had focused on the AoI targeted by the test item claiming a strong cognitive validity. Furthermore, these findings matched self-report of the participants in retrospective questionnaires to some extent. Specifically, the participants had been 68.4% accurate in their self-report, showing that majority of the self-assessment was accurate. However, they had been 31.6% inaccurate in their self reports. Bax and Weir (2012) argue that this creates doubt upon studies that depend heavily on retrospective reports to gain insights about cognitive processing. In brief, by mentioning the limitations of traditional retrospective methods to test validation and to uncover cognitive processing, this study demonstrated that eye-tracking is a valuable tool to assist in the validation of test items, to understand cognitive processes of readers under test conditions, and to shed light onto the ways in which test items can perform when eliciting particular cognitive processes in reading (Bax & Weir, 2012).

Bax (2013) conducted a study to investigate test takers' cognitive processing while taking on-screen IELTS (International English Language Testing System) reading test items in order to evaluate the cognitive validity of these and similar reading test items. Bax focused on differences in reading behaviors of successful and unsuccessful participants while completing the IELTS reading test items. Seventy-one Malaysian undergraduates, whose L1 includes Tamil, Chinese and others, took an onscreen test consisting of two IELTS reading passages with 11 related questions in given 30 minutes. The passages were split across three pages while questions for each passage remained on the same page to avoid scrolling. EMs of a random sample of the participants (n = 38) were recorded through eye-tracking to investigate "careful local reading" and "expeditious local reading" (Khalifa & Weir, 2009). A sentence completion task was used to assess careful local reading while a matching task was used to assess expeditious local reading. In addition, stimulated recall interview data were collected from a sample of participants (n = 20) to triangulate and assist the interpretation of the eye-tracking data. For the analysis, the EMs of the students whose answer for the item correct were compared to those of other students whose answer was wrong for that item to see whether differences in EMs could help explaining success or failure. Considering EMs and stimulated recall interview data, results of the analysis were statistically significant for five eleventh of items, so these five items were focused on to reach a conclusion. With respect to cognitive validity, the findings of the study showed the potential of eye-tracking to investigate cognitive validity in reading tests as it was seen that the EMs of the successful and unsuccessful readers differed. More specifically, the findings illustrated that the successful participants used the kinds of cognitive strategies that would be expected in real-life academic situations while the unsuccessful participants did not on many items. For instance, the unsuccessful students spent on average 163.25 seconds on the correct page of the text while the successful students on average spent 102.55 seconds on larger areas of the texts. Also, the successful students had significantly more fixations on the correct paragraph than the unsuccessful participants. These results showed that the unsuccessful participants were not able to read expeditiously to find the location of the answer which led spending significantly more time looking (in vain) than the successful participants. And, this was supported by recall interviews too, as the successful participants reported that they used conscious metacognitive strategies to read expeditiously; however, the unsuccessful participants reported no such conscious strategies, but "seemed rather to be searching almost at random, and with no strategic purpose" (Bax, 2013, p. 460) showing expeditious reading is connected to metacognitive awareness and differentiates successful readers from unsuccessful readers. In short, Bax (2013) concluded that eye-tracking analysis is valuable to evaluate the cognitive processing of test takers in language tests, and in such tests, proficient readers' EM behaviors are significantly different from those of unsuccessful readers because, most probably, of their behavior linked to different cognitive processing at different levels of comprehension.

Solheim and Uppstad (2011) designed a study to address the need for methodological reflection on how to validate inferences made on the basis of test scores. The EM data of 18 seventh graders were analyzed for the study.
The participants read a text in their L1 by being aware of the fact that they would answer questions related to it afterwards, and they were given unlimited time to study the text. Then, they were given a question sheet including a mixture of MCQs and constructed response items with unlimited time to answer them. The participants had access to the text and were allowed to look back at it when answering the questions. The data were analyzed in three parts. In Part 1, the relationship between latent response times and outcome scores was investigated and no significant relationship was observed between how much time the participants spent on reading the text or on answering the questions and how well they scored on the items. In Part 2, the focus was on the extent to which the participants answering the question correctly exhibited a reading behavior different from that of the participants who did not with the use of gaze duration, TRT on different sections of the passage, and integrative saccades (i.e., transactions between semantically related segments of verbal text and illustration). It was seen that some participants divided their attention equally between the text and the illustration, but got different comprehension scores, meaning there was not a significant difference between the reading behaviors of the readers who answered the item correctly and those who did not. In Part 3, the characteristics of reading behavior were qualitatively investigated by comparing the first reading (global purpose reading) with reading while answering the question (task oriented reading). The participants were categorized into four groups as "task-oriented readers" (readers who scored the item correctly and who read the relevant parts of the verbal text while answering the question), "effortful readers" (readers

who scored the item incorrectly but who read the relevant parts of the verbal text while answering the question), “first-time readers” (readers who scored the item correctly but who did not read the relevant parts of the verbal text but focused on the illustration while answering the question), and “non-strategic readers” (readers who scored the item incorrectly and who did not read the relevant parts of the verbal text but focused on the illustration while answering the question). When the groups were compared, it was observed that the first-time readers spent more time on the first reading of the text (average of 138 seconds) than the non-strategic readers (average of 108 seconds). Also, the first-time readers were more likely to adjust their reading speed to the difficulty of the text and they spent 50.4% of the TRT in text relevant area while the non-strategic readers spent 40%. The task-oriented readers, on the other hand, used slightly more time to answer the question, and spent a larger percent of their time on the text passage (37.6% versus 18.3%) compared to other groups. With respect to the first-time readers, the task-oriented readers spent less time on their first reading, they spent less time on the TR area (43.5% versus 50.4%), and they had fewer integrative saccades (1.5 versus 5.0). The final group, effortful readers, spent most time of all on answering the question. In brief, it was found out that the task-oriented readers were the most successful of all on the test (total score of 7.5 points). The first-time readers, first, read the text carefully and trusted in their first reading while answering the questions while the task-oriented readers skimmed the text first and let the individual questions direct their future reading. The non-strategic readers were the least successful of all (total score of 4.3 points) by displaying a limited range of reading strategies and using the least TRT to complete the task. The effortful readers, although they knew necessary reading strategies, could not use them effectively to succeed in the task (total score of 4.8 points). Based on the results, Solheim and Uppstad (2011) argue that eye-tracking is a useful tool to study test validation by displaying a more detailed picture of cognitive processes of readers.

In conclusion, eye-tracking technology is used by researchers because of its advantages over other on-line and off-line methods to investigate the cognitive processes in reading and the cognitive validity of reading test items (Dussias, 2010; Hyönä & Lorch, 2004; Solheim & Uppstad, 2011), and study results (Bax & Weir, 2012; Bax, 2013; Solheim & Uppstad, 2011) showed the feasibility of using eye-tracking in cognitive validity research.

3.10 Significance of the study

When the condition that cognitive processes which a test elicits from a reader resemble the processes which a test taker would use in non-test conditions is provided, the test is accepted as valid. In non-test conditions, reading comprehension necessitates readers to go through levels of comprehension mixing lower-level skills with higher-level skills. Thus, a reading test prepared to assess academic language proficiency of readers should require readers to go through these literal and inferential levels of comprehension to be cognitively valid.

At that point, the importance of careful examination of the cognitive processes of readers when they are taking an academic reading test manifests itself. However, despite EM research on strategy use to understand cognitive processes in reading (Carpenter & Just, 1989; Kaakinen & Hyönä, 2007), not much is known about the reading behavior of the participants when they read longer texts to fulfill a reading comprehension task that includes literal and inferential question types that involve different levels of cognitive load. Hence, whether EM behaviors of individuals answering questions, whose nature is different (e.g., literal or inferential), correctly and incorrectly show any differences is piquant, especially in the case of L2 learners.

In that vein, first, this study was motivated to determine if EM behaviors in different parts of the text and questions differed according to the participants' performance, i.e. those who answered a question correctly vs. those who provided an incorrect answer, and the type of question, i.e. literal vs. inferential. Second, it aimed to examine the relationship between reading comprehension and EM behaviors in different parts of the text and questions.

In sum, this study is significant in terms of its findings for two reasons. First, it sheds light on the issue of whether eye movements could provide valuable information about test takers' cognitive processes as they answer literal and inferential questions in a standardized reading test. Second, there are only a handful of studies in educational research examining cognitive processing and cognitive validity in L2 through eye-tracking, so the results will enrich the educational research and inspire further research in the area.

CHAPTER 4

METHODOLOGY

The methodology section presents the methods and procedures that were used in the present study. Research questions and hypotheses related to these questions are stated. Following the hypotheses, participants’ characteristics and the instruments used for data collection are provided in detail. Moreover, sample screenshots are displayed under relevant sections in order to elucidate materials lucidly. The chapter ends with a detailed explanation of how the data were analyzed.

4.1 Research questions and hypotheses

The main purpose of the current study was to explore the EM behaviors of L2 readers answering literal and inferential questions with a view to understand whether EM data could provide any insights regarding how the items function. Specifically, the first aim was to determine whether EM behaviors in different parts of the text and questions differed according to participants' performance (i.e., those who answered a question correctly vs. those who provided an incorrect answer) and the type of question (literal vs. inferential). A second aim of the study was to investigate the relationship between reading comprehension and EM behaviors in different parts of the text and questions. The following research questions were investigated:

1. What are the psychometric properties of the NDRT in terms of the test's internal consistency reliability, item difficulty, and item analysis?
2. Are there any differences between test takers answering a question correctly and those incorrectly in terms of their EMs (FPRT, TFC, and TRT) in five AoIs: part of the text that is relevant to the question (TR), part of the text that is

irrelevant to the question (TIR), question stem (QS), correct option (CO), incorrect option (INCO)?

3. Do EMs (FPRT, TFC, and TRT) in five AoIs (TR, TIR, QS, CO, and INCO) differ when the test takers are engaged in answering literal versus inferential questions?
4. Based on the first 10 fixations, do the test takers answering a question correctly and those incorrectly differ in the AoIs they focus on while answering literal and inferential questions?
5. Do reading comprehension scores correlate with EMs (FPRT, TFC, and TRT) in the following AoIs: TR, TIR, QS, CO, and INCO?

First, based on previous research (Bax, 2013; Bax & Weir, 2012; Solheim & Uppstad, 2011), it was predicted that the participants answering a question correctly would have longer FPRT and TRT with more frequent fixations in the TR, QS and CO AoIs but they would have shorter FPRT, TRT, and less frequent fixations in the TIR and INCO AoIs compared to those participants answering the question incorrectly (Hypothesis 1). Second, it was hypothesized that the participants' fixations would be longer and more frequent in all of the AoIs while answering literal questions compared to inferential questions (Hypothesis 2) since literal questions are more text-bound and their answers can be directly located in the text (Kintsch, 1998). Third, it was anticipated that both the correct and incorrect responders would first focus on the QS AoI and the options of the question before referring to the text (Bax & Weir, 2012; Nevo, 1989) (Hypothesis 3). Finally, it was hypothesized that reading comprehension scores would correlate positively with EMs in the TR, QS, and CO AoIs but negatively with EMs in the TIR and INCO AoIs (Bax, 2013; Bax & Weir, 2012; Kintsch, 1998) (Hypothesis 4).

4.2 Participants

The data were collected from 33 (26 female and 7 male) freshman students majoring in English Language Teacher Education (ELT) in the Foreign Language Education Department at Boğaziçi University, where the medium of instruction is English. Since one participant's EMs could not be recorded properly due to calibration problems, his data were excluded from the analyses. Thus, the data obtained from a total of 32 participants (26 female and 6 male) were included in the analyses. All of the participants had normal or corrected-to-normal (with soft contact lenses) vision.

The sample was rather homogeneous in terms of their L1 and educational background. All of the participants were native speakers of Turkish. Their ages ranged from 18 to 19. Their years of exposure to English, albeit limited to formal learning settings, ranged from 8 to 10 years. For that reason, it can be said that they started to learn L2 English in middle childhood when they were fourth graders. Because of the fact that individuals who are exposed to L2 “in middle childhood (around 8-10 years) or later” are defined as “late L2 learners” (van Hell & Tokowicz, 2010, p. 44), the participants of the current study can be classified as late L2 learners.

Having passed a National Foreign Language Test (YDS) in English 3 to 8 months before becoming a regular ELT student at Boğaziçi University, the participants were used to reading both short and long passages as well as answering literal and inferential questions related to texts. As such, the participants were familiar with reading texts and answering different question types in L2 English in pen and paper as well as on screen. Also, all the participants had passed the Boğaziçi University English Proficiency Test (BUEPT) and were accepted to their programs as regular ELT freshman students. It should be noted that the minimum pass mark on the BUEPT is equal to 550 on the paper-based TOEFL, 79 on TOEFL IBT, and 6.5

on IELTS Academic. Thus, they were considered to be relatively advanced-level learners of English.

Furthermore, by the time data collection started, all of the participants had taken an “Introduction to Computers” undergraduate course through which they learnt how to use computer operations and concepts, word processing, graphics, multimedia, and databases. Thus, it is possible to say that they were computer literate.

No course credit or any kind of reward to externally motivate the students was given for their participation. The students volunteered to take part in the study and all of them signed appropriate consent forms prepared in L1 Turkish.

4.3 Instruments

The reading comprehension component of the Nelson-Denny Reading Test (NDRT) (Form G; Brown, Fishco, & Hanna, 1993) was used to collect quantitative data. This test was implemented to gather data regarding the participants' EMs as they answered reading comprehension questions in L2.

4.3.1 Reading comprehension test

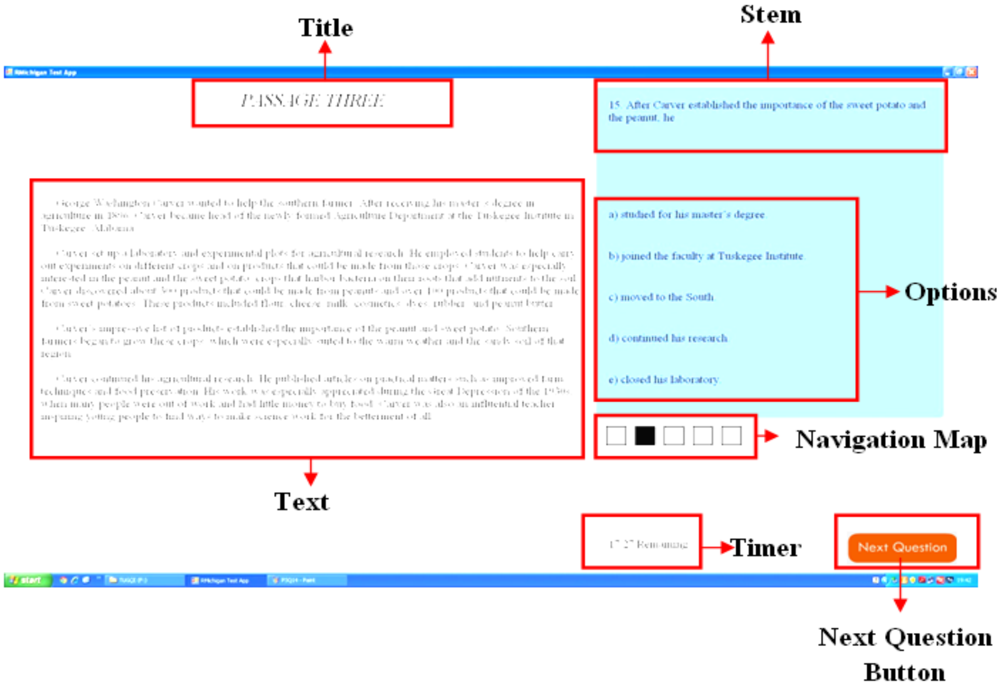
The NDRT (Form G; Brown, et al., 1993) excluding its vocabulary subtest was used as a comprehension test. This standardized test aiming to measure vocabulary, comprehension, and reading rate is developed for L1 readers; however, the test developers suggest its use with students of English as a second or foreign language by giving extended time. The NDRT includes seven passages. The first passage, which is the longest with 603 words, has eight questions related to it. The length of the other six passages ranges from 156 to 237 words and each passage has related five questions. In total, there are 38 MCQs with five options.

Of the 38 questions, 21 were literal questions (question 1, 2, 3, 4, 6, 9, 10, 12, 14, 15, 16, 19, 20, 24, 25, 26, 29, 30, 31, 34, 35) while 17 were inferential questions (question 5, 7, 8, 11, 13, 17, 18, 21, 22, 23, 27, 28, 32, 33, 36, 37, 38) based on two researchers' separate categorization and their consensus. A question was classified as literal if its answer could be provided using information explicitly stated within the text; and inferential if its answer had to be inferred either based on textual information and the integration of information located in the text with the readers' world knowledge. However, it should be noted that inferential questions in the NDRT necessitated making more of local inferencing rather than elaborative or global inferencing. Thus, inferential questions were expected to assess readers' understanding of relationships between multiple ideas located in the text with the ones that were dislocated by tapping the higher level of comprehension. As such, these two types of questions were to assess different levels of comprehension. Also, it is necessary to note that because there were no evaluative questions in the NDRT, the test was not appropriate to analyze the third comprehension level, evaluative comprehension.

In this study, the original NDRT was reproduced in electronic format without any semantic, syntactic or vocabulary adjustments. However, the layout of the pages was changed as each page displayed a passage, on the left, and its related question, on the right (see Figure 1). Additionally, a timer showed the remaining time (30 minutes were allotted to finish the test) and a navigation map showed where a participant was. Participants had to answer the question on the page in order to see the next question, and it was not possible to go back to already-answered questions.

After choosing an option, the participants needed to click on the "Next Question" button to see the next questions.

After piloting the electronic version of the NDRT with five (three female and two male) senior students studying ELT at Boğaziçi University and making necessary changes, the page layout took its final form (see Figure 1). Administering the NDRT in electronic format allowed storing the responses to questions in a folder so that each participant's correct and incorrect answers could be recorded. The maximum total score on the test was 38, as one-point was allotted for each correct answer.



PASSAGE THREE

George Washington Carver wanted to help the southern farmer. After receiving his master's degree in agriculture in 1896, Carver became head of the newly formed Agriculture Department at the Tuskegee Institute in Tuskegee, Alabama.

Carver set up a laboratory and experimental plots for agricultural research. He employed students to help carry out experiments on different crops and on products that could be made from those crops. Carver was especially interested in the peanut and the sweet potato, crops that harbor bacteria on their roots that add nutrients to the soil. Carver discovered about 300 products that could be made from peanuts and over 100 products that could be made from sweet potatoes. These products included flour, cheese, milk, cosmetics, dyes, rubber, and peanut butter.

Carver's impressive list of products established the importance of the peanut and sweet potato. Southern farmers began to grow these crops, which were especially suited to the warm weather and the sandy soil of that region.

Carver continued his agricultural research. He published articles on practical matters such as improved farm techniques and food preservation. His work was especially appreciated during the Great Depression of the 1930s, when many people were out of work and had little money to buy food. Carver was also an influential teacher, inspiring young people to find ways to make science work for the betterment of all.

15. After Carver established the importance of the sweet potato and the peanut, he

a) studied for his master's degree.

b) joined the faculty at Tuskegee Institute.

c) moved to the South.

d) continued his research.

e) closed his laboratory.

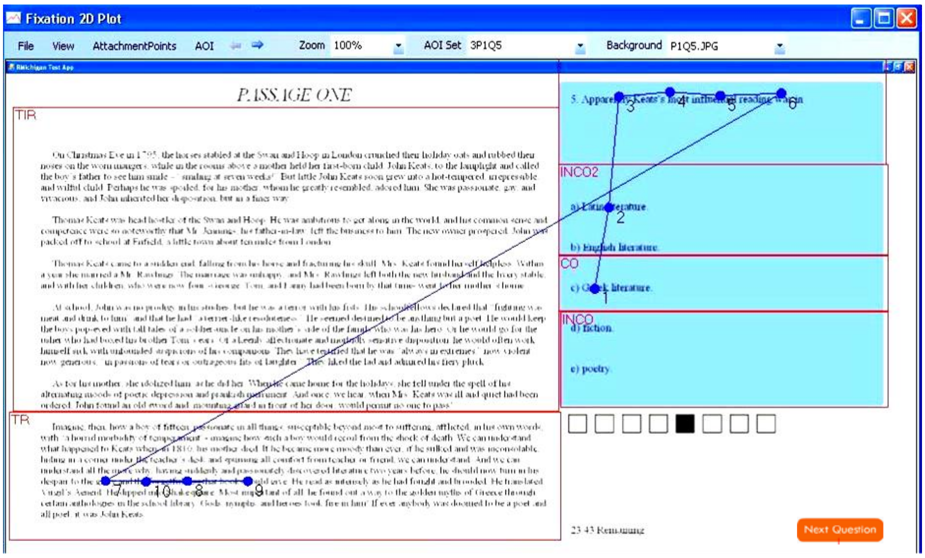
□ ■ □ □ □

Fig. 1. A screenshot displaying page layout.

4.3.2 EMs

In order to obtain process-based data regarding whether any differences are observed between the behaviors of the participants answering a question correctly and incorrectly, EM data were collected (Bax & Weir, 2012). Thus, the concurrent "real-time” measures of the FPRT, TFC and TRT were calculated (Hyönä & Lorch, 2004; Roberts & Siyanova-Chanturia, 2013) for five different targets (TR, TIR, QS, CO, INCO) for each question on the NDRT. And also, the places of the first 10 fixations were recorded for each question on the NDRT.

The first calculated index was the FPRT, the duration of the first fixation within an AoI, which was determined considering a suitable margin of error to allow for individual variation in fixation and taking account of typical saccade lengths for each measurement, regardless of whether it is the only fixation or the first of several fixations within the region (see Figure 2). However, when an AoI gets larger, the possibility of additional fixations on this area increases as well (Roberts & Siyanova-Chanturia, 2013). Considering the FPRT may not be informative enough for an area larger than a single word, the TRT and the TFC were also measured.



Zoom 100%

AOI Set 3P1Q5

Background P1Q5.JPG

PASSAGE ONE

On Christmas Eve in 1795, the horses stabled at the Swan and Hoop in London crunched their holiday oats and rubbed their noses on the worn mangers, while in the rooms above a mother held her first-born child, John Keats, to the lamplight and called the boy's father to see him smile – "smiling at seven weeks!" But little John Keats soon grew into a hot-tempered, irrepressible and willful child. Perhaps he was spoiled, for his mother, whom he greatly resembled, adored him. She was passionate, gay, and vivacious, and John inherited her disposition, but in a finer way.

Thomas Keats was head hostler of the Swan and Hoop. He was ambitious to get along in the world, and his common sense and competence were so noteworthy that Mr. Jennings, his father-in-law, left the business to him. The new owner prospered. John was packed off to school at Enfield, a little town about ten miles from London.

Thomas Keats came to a sudden end, falling from his horse and fracturing his skull. Mrs. Keats found herself helpless. Within a year she married a Mr. Rawlings. The marriage was unhappy, and Mrs. Rawlings left both the new husband and the livery stable, and with her children, who were now four – George, Tom, and Fanny had been born by that time– went to her mother's home.

At school, John was no prodigy in his studies, but he was a terror with his fists. His schoolfellows declared that "fighting was meat and drink to him" and that he had "a terrier-like resoluteness." He seemed destined to be anything but a poet. He would keep the boys pop-eyed with tall tales of a soldier-uncle on his mother's side of the family who was his hero. Or he would go for the usher who had boxed his brother Tom's ears. Of a keenly affectionate and morbidly sensitive disposition, he would often work himself sick with unfounded suspicions of his companions. They have testified that he was "always in extremes" now violent now generous, "in paroxysms of tears or outrageous fits of laughter." They liked the lad and admired his fiery pluck.

As for his mother, she idolized him as he did her. When he came home for the holidays, she fell under the spell of his alternating moods of poetic depression and prankish merriment. And once, we hear, when Mrs. Keats was ill and quiet had been ordered, John found an old sword and, mounting guard in front of her door, would permit no one to pass.'

Imagine, then, how a boy of fifteen, passionate in all things, susceptible beyond most to suffering, afflicted, in his own words, with "a horrid morbidity of temperament" – imagine how such a boy would recoil from the shock of death. We can understand what happened to Keats when, in 1810, his mother died. If he became more moody than ever, if he sulked and was inconsolable, hiding in a corner under the teacher's desk, and spurning all comfort from teacher or friend, we can understand. And we can understand all the more why, having suddenly and passionately discovered literature two years before, he should now turn in his despair to the [illegible] and [illegible] that books could give. He read as intensely as he had fought and brooded. He translated Virgil's Aeneid. He dipped into Shakespeare. Most important of all, he found out a way to the golden myths of Greece through certain anthologies in the school library. Gods, nymphs, and heroes took fire in him! If ever anybody was doomed to be a poet and all poet, it was John Keats.

5. Apparently Keats's most influential reading was in

a) Latin literature.

b) English literature.

c) Greek literature.

d) fiction.

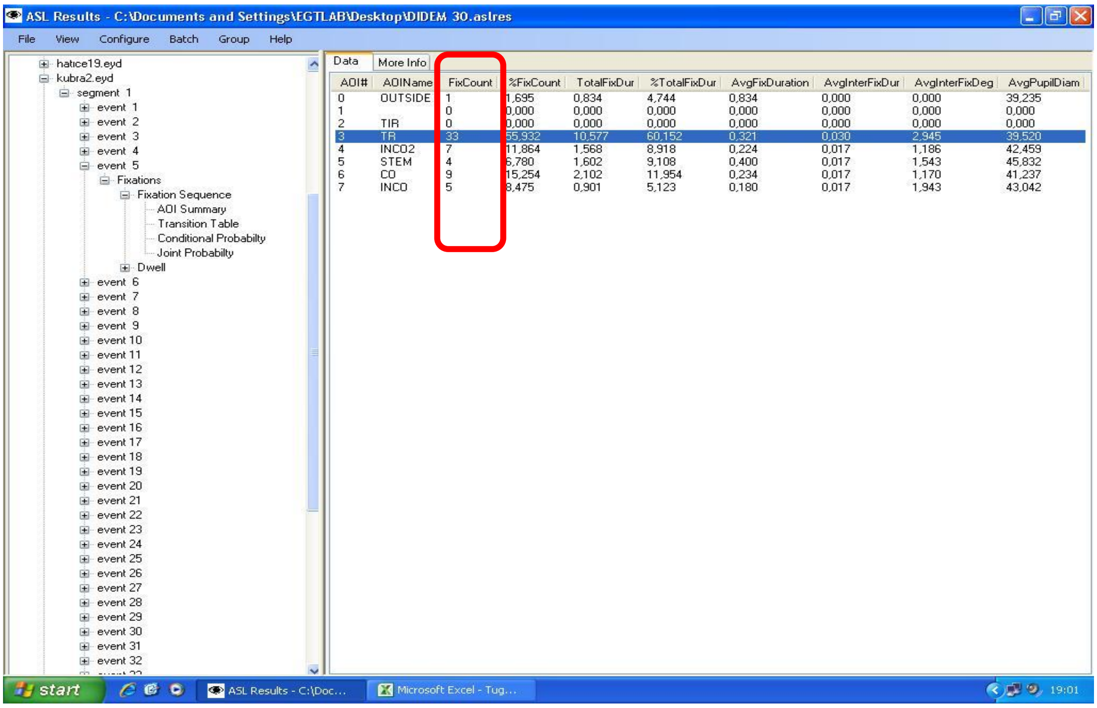
e) poetry.

□ □ □ □ ■ □ □ □

23:43 Remaining

Fig. 2. A screenshot displaying the FPRT.

The TFC (see Figure 3), which is the number of all fixations made within a given AoI (Roberts & Siyanova-Chanturia, 2013), was not a measure of processing time, rather it was calculated to reveal how many times the target was fixated.

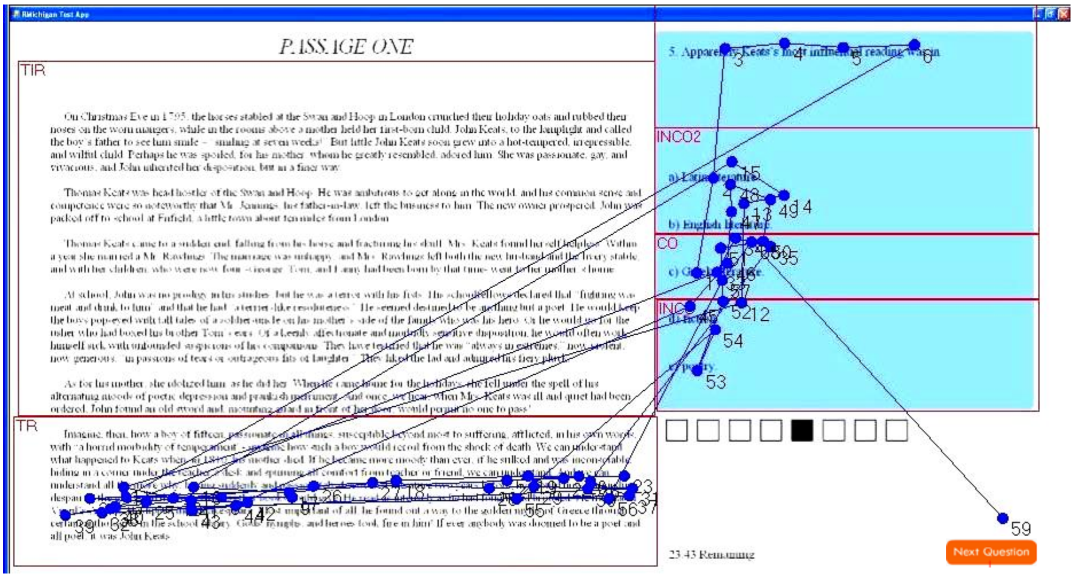


AOI#	AOIName	FixCount	%FixCount	TotalFixDur	%TotalFixDur	AvgFixDuration	AvgInterFixDur	AvgInterFixDeg	AvgPupilDiam
0	OUTSIDE	1	1,695	0,834	4,744	0,834	0,000	0,000	39,235
1		0	0,000	0,000	0,000	0,000	0,000	0,000	0,000
2	TIR	0	0,000	0,000	0,000	0,000	0,000	0,000	0,000
3	TR	33	55,932	10,577	60,152	0,321	0,030	2,945	39,520
4	INCO2	7	11,864	1,568	8,918	0,224	0,017	1,186	42,459
5	STEM	4	6,780	1,602	9,108	0,400	0,017	1,543	45,832
6	CO	9	15,254	2,102	11,954	0,234	0,017	1,170	41,237
7	INCO	5	8,475	0,901	5,123	0,180	0,017	1,943	43,042

Fig. 3. A screenshot displaying the TFC.

The third index, the TRT measure, consists of all fixations that landed on the target (see Figure 4), which is a single word or a longer phrase, and reveals the total amount of time the participant spent on a given target (Roberts & Siyanova-Chanturia, 2013; Siyanova-Chanturia, Conklin, & Schmitt, 2011; Siyanova-Chanturia, Conklin, & van Heuven, 2011). The TRT was measured considering the fact that when an AoI is larger than a single-word, it is better to calculate the TRT as a major eye-movement measure (Rayner, 1998) as it is a combination of initial processing time as well as the time that may have been spent recovering from processing difficulties (Liversedge, Paterson, & Pickering 1998).

Finally, the sequence of the first 10 fixations was listed for each question on the NDRT (see Figure 5). An analysis of the sequence of the first 10 fixations shows which AoIs are fixated in which order. For that reason, it is useful to observe the participants’ EM behaviors and reading strategies (Hannus & Hyönä, 1999).



RMichigan Test App

PASSAGE ONE

On Christmas Eve in 1795, the horses stabled at the Swan and Hoop in London crunched their holiday oats and rubbed their noses on the worn mangers, while in the rooms above a mother held her first-born child. John Keats, to the lamplight and called the boy's father to see him smile – "smiling at seven weeks!" But little John Keats soon grew into a hot-tempered, irrepressible, and wilful child. Perhaps he was spoiled, for his mother, whom he greatly resembled, adored him. She was passionate, gay, and vivacious, and John inherited her disposition, but in a finer way.

Thomas Keats was head hostler of the Swan and Hoop. He was ambitious to get along in the world, and his common sense and competence were so noteworthy that Mr. Jennings, his father-in-law, left the business to him. The new owner prospered. John was packed off to school at Enfield, a little town about ten miles from London.

Thomas Keats came to a sudden end, falling from his horse and fracturing his skull. Mrs. Keats found herself helpless. Within a year she married a Mr. Rawlings. The marriage was unhappy, and Mrs. Rawlings left both the new husband and the livery stable, and with her children, who were now four – George, Tom, and Fanny had been born by that time – went to her mother's home.

At school, John was no prodigy in his studies, but he was a terror with his fists. His schoolfellows declared that "fighting was meat and drink to him" and that he had "a terrier-like resoluteness." He seemed destined to be anything but a poet. He would keep the boys pop-eyed with tall tales of a soldier-uncle on his mother's side of the family who was his hero. Or he would go for the usher who had boxed his brother Tom's ears. Of a keenly affectionate and morbidly sensitive disposition, he would often work himself sick with unfounded suspicions of his companions. They have testified that he was "always in extremes," now violent, now generous, "in passions of tears or outrageous fits of laughter." They liked the lad and admired his fiery pluck.

As for his mother, she idolized him as he did her. When he came home for the holidays, she fell under the spell of his alternating moods of poetic depression and prankish merriment. And once, we hear, when Mrs. Keats was ill and quiet had been ordered, John found an old sword and, mounting guard in front of her door, would permit no one to pass.

Imagine, then, how a boy of fifteen, passionate in all things, susceptible beyond most to suffering, afflicted, in his own words, with "a horrid morbidity of temperament," – imagine how such a boy would recoil from the shock of death. We can understand what happened to Keats when, in 1810, his mother died. If he became more moody than ever, if he sulked and was inconsolable, hiding in a corner under the teacher's desk and spurning all comfort from teacher or friend, we can understand. And we can understand all [illegible] despair [illegible] [illegible] most important of all, he found out a way to the golden myths of Greece through certain authors in the school library. Gods, nymphs, and heroes took fire in him! If ever anybody was doomed to be a poet and all poet, it was John Keats.

5. Apparently Keats's most influential reading was in

a) Latin literature.

b) English literature.

c) Greek literature.

d) fiction.

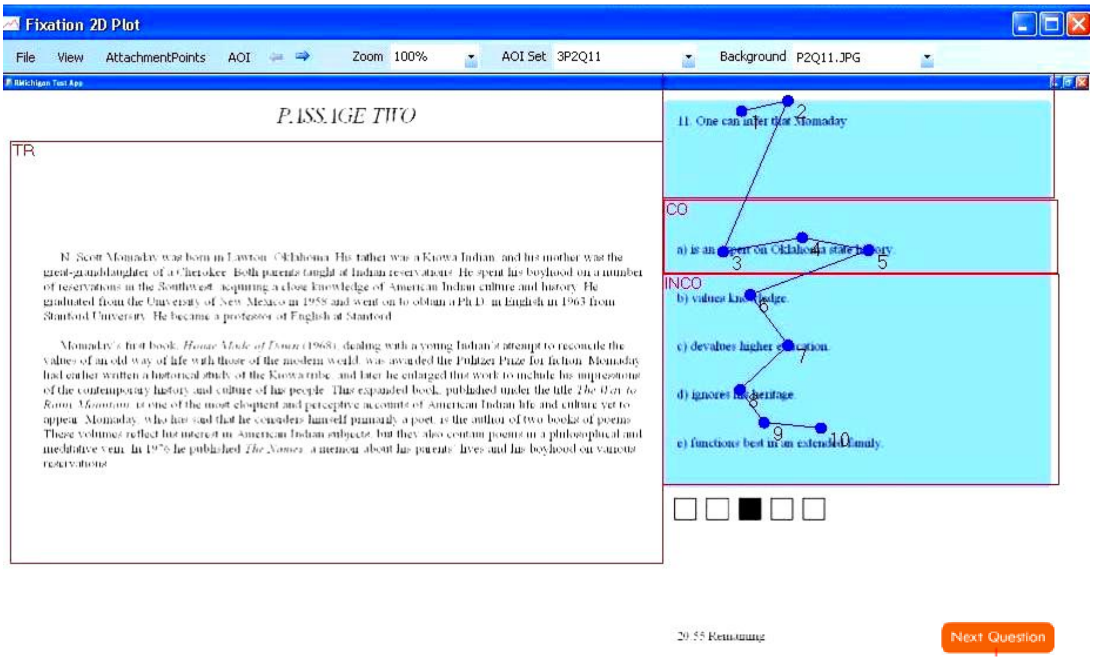
e) poetry.

□ □ □ □ ■ □ □ □

23:43 Remaining

Next Question

Fig. 4. A screenshot displaying the TRT.



Fixation 2D Plot

Zoom 100%   AOI Set 3P2Q11   Background P2Q11.JPG

RMichigan Test App

PASSAGE TWO

N. Scott Momaday was born in Lawton, Oklahoma. His father was a Kiowa Indian, and his mother was the great-granddaughter of a Cherokee. Both parents taught at Indian reservations. He spent his boyhood on a number of reservations in the Southwest, acquiring a close knowledge of American Indian culture and history. He graduated from the University of New Mexico in 1958 and went on to obtain a Ph.D. in English in 1963 from Stanford University. He became a professor of English at Stanford.

Momaday's first book, *House Made of Dawn* (1968), dealing with a young Indian's attempt to reconcile the values of an old way of life with those of the modern world, was awarded the Pulitzer Prize for fiction. Momaday had earlier written a historical study of the Kiowa tribe, and later he enlarged this work to include his impressions of the contemporary history and culture of his people. This expanded book, published under the title *The Way to Rainy Mountain*, is one of the most eloquent and perceptive accounts of American Indian life and culture yet to appear. Momaday, who has said that he considers himself primarily a poet, is the author of two books of poems. These volumes reflect his interest in American Indian subjects, but they also contain poems in a philosophical and meditative vein. In 1976 he published *The Names*, a memoir about his parents' lives and his boyhood on various reservations.

11. One can infer that Momaday

a) is an expert on Oklahoma state history.

b) values knowledge.

c) devalues higher education.

d) ignores his heritage.

e) functions best in an extended family.

□ □ ■ □ □

29:55 Remaining

Next Question

Fig. 5. A screenshot displaying the sequence of the first 10 fixations.

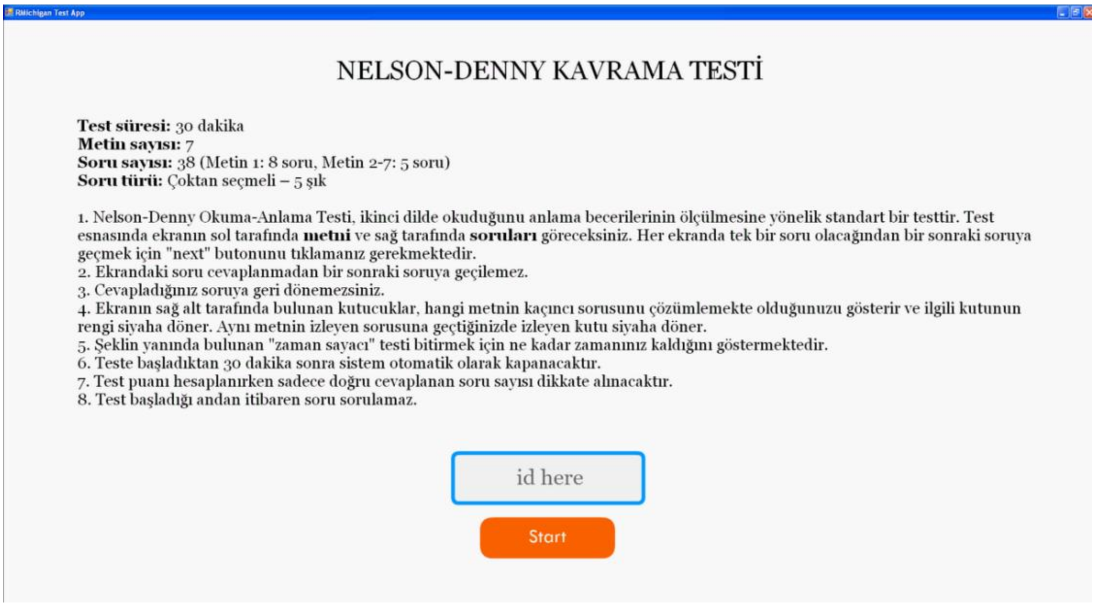
To track the participants' EMs, the Applied Science Laboratories' (ASL) D6 Desk mounted Optics remote eye tracker was used in the study. The D6 Optics module, which was on a stand under a 19-inch subject display monitor, consisted of an eye-tracking camera and a head-tracking camera. The eye-tracking camera was directly connected to the interface PC and located on the left of the module while the head-tracking camera was placed on the right, hence, it was not necessary to use a head-mounted eye-tracking with a chin rest or bite bar.

4.4 Procedures

First, the participants were informed about the general outline of the study and the electronic version of the NDRT. However, the main purpose of the experiment was kept vague. The whole procedure consisted of online (electronic version of the NDRT) data collection session.

The online study was conducted in Boğaziçi University's fluorescent-lit and noise insulted eye-tracking laboratory where the desktop and mounted eye-tracking system (ASL EYE-TRAC®6) is set up. Each participant was seated with eyes at a distance of 60cm from the screen. Before the calibration process, the participants watched a demo video (1 minute 43 seconds) to have an idea about what they were expected to do, and the video started with the NDRT instruction page, which was in L1 Turkish (see Figure 6). The instruction page included crucial information on page layout, number of the passages, number and format of the questions, time given to complete the test and the meaning of the symbols and buttons seen on each page.

They were also informed about score calculation procedure via the demo video, and notified that any head or body movement might distort the unobtrusive sensitivity of eye-tracking. Following the essential instruction and warning session, conventional nine-point calibration process was completed and the participants started to take the electronic version of the NDRT. While they were taking the test, real time online eye and head movement data were collected. The whole online data collection process lasted for approximately 30 minutes in each individual session excluding the time spent on calibration and demo, and the data collection procedure ended after tracking eye data of 33 participants.



NELSON-DENNY KAVRAMA TESTİ

Test süresi: 30 dakika
Metin sayısı: 7
Soru sayısı: 38 (Metin 1: 8 soru, Metin 2-7: 5 soru)
Soru türü: Çoktan seçmeli – 5 şık

1. Nelson-Denny Okuma-Anlama Testi, ikinci dilde okuduğunu anlama becerilerinin ölçülmesine yönelik standart bir testtir. Test esnasında ekranın sol tarafında **metni** ve sağ tarafında **soruları** göreceksiniz. Her ekranda tek bir soru olacağından bir sonraki soruya geçmek için "next" butonunu tıklamanız gerekmektedir.
2. Ekrandaki soru cevaplanmadan bir sonraki soruya geçilemez.
3. Cevapladığınız soruya geri dönemezsiniz.
4. Ekranın sağ alt tarafında bulunan kutucuklar, hangi metnin kaçıncı sorusunu çözümlemekte olduğunuzu gösterir ve ilgili kutunun rengi siyaha döner. Aynı metnin izleyen sorusuna geçtiğinizde izleyen kutu siyaha döner.
5. Şeklin yanında bulunan "zaman sayacı" testi bitirmek için ne kadar zamanınız kaldığını göstermektedir.
6. Teste başladıktan 30 dakika sonra sistem otomatik olarak kapanacaktır.
7. Test puanı hesaplanırken sadece doğru cevaplanan soru sayısı dikkate alınacaktır.
8. Test başladığı andan itibaren soru sorulamaz.

Fig. 6. A screenshot displaying the NDRT instructions in L1 Turkish.

4.5 Data analyses

Each correct answer on the NDRT was given one point while no points were given to each incorrect answer. Thus, the maximum possible score for the NDRT was 38. An item analysis was carried out in order to investigate the difficulty level and discrimination power of the items. The internal consistency reliability of the test was assessed through Cronbach's alpha.

As measures of EMs, the average FPRT, TFC and TRT on the TR, TIR, QS, CO, and INCO AoIs for selected literal and inferential questions were calculated for the participants answering a question correctly and incorrectly separately. Only those items for which there were at least five correct and incorrect responses were included for this analysis as it was thought that reliable comparisons would not be possible with fewer numbers of responses.

In order to determine whether EMs differed between literal and inferential questions, the proportions of time (FPRT and TRT) spent by the participants while looking at the different areas were compared. Similarly, the proportions of total frequency counts were compared between the two types of questions. The reason why the proportions were analyzed instead of actual EM data was due to the differences in size across the five regions. Once the proportions were obtained, they were compared through a 2 x 5 repeated measures ANOVA with type of question (literal, inferential) and AoI (TR, TIR, QS, CO, INCO) as within group factors.

The correlations of reading scores with EMs (FPRT, TFC, TRT) across the AoIs were obtained in order to see the relationships between reading performance and EM measures. Also, a 2 x 5 repeated measures ANOVA with type of question (literal, inferential) and AoI (TR, TIR, QS, CO, INCO) as within group factors was conducted to be able to see the effects of question type on the FPRT, TFC and TRT separately.

Finally, for each of the chosen 11 questions on the NDRT, the sequence of the first 10 fixations was listed for further investigation to complement the results of the FPRT, TFC, and TRT.

CHAPTER 5

RESULTS

5.1 Item characteristics of the NDRT

An item analysis was carried out to examine the reliability of the test as well as the difficulty and discrimination characteristics of the items (see Appendix A). The Cronbach's Alpha was .43 for the total of 38 items, .27 for the literal items, and .24 for the inferential items. These alpha coefficients are not at desired levels. An examination of the characteristics was made in order so as to get a further understanding of low alpha coefficients.

The items seem to be relatively easy for the participants, as the average item difficulty was 79.86. Item total correlations indicate that the items do not have sufficient discrimination power either (average item total correlation = .09). Low level of item discrimination is partly due to the fact that the test was too easy for the participants. Hence, these findings suggest that the NDRT items do not sufficiently differentiate between students with higher and lower reading scores in the sample.

The next section is a comparison of the EM behaviors of the participants answering the selected question correctly and incorrectly to explore if they could provide further insights about how the items function for the sample of the study.

5.2 Comparing correct and incorrect responders' EMs on selected items

5.2.1 The FPRT

Descriptive statistics for the FPRT for literal questions (see Appendix B) and inferential questions (see Appendix C) were examined in detail, and an examination of the FPRT for literal questions indicated that the correct responders had longer fixations in the TR AoI of Question 34, 19, and 3, the QS AoI of Question 6, 19, 3, and 14, and the CO AoI of Question 6, 31, 19, and 14, as expected. In addition, they had shorter fixations in the TIR AoI of Question 34, 31, 3, and 14 and the INCO AoI of Question 6, 19, and 3.

As for the FPRT for inferential questions (see Appendix C), the correct responders had longer fixations in the TR AoI of Question 28, 33, 13, 36, the QS AoI of Question 33, 13 and 36, and the CO AoI of Question 28, 33, 13. Additionally, when the fixations landed in the TIR AoI were analyzed, it was observed that the correct responders' FPRT was shorter than the incorrect responders for Question 36, and 17. They also had shorter fixations in the INCO AoI of Question 13, 36, 17.

Table 1 provides a summary of the analyses related to the FPRT for literal and inferential questions.

Table 1. The Summary Table for the FPRT

	AoIs				
	TR	TIR	QS	CO	INCO
Literal Questions					
Correct Responders	3	2	4	4	3
Incorrect Responders	3	4	2	2	3
Inferential Questions					
Correct Responders	4	3	3	3	2
Incorrect Responders	1	2	2	2	3
Total					
Correct Responders	7	5	7	7	5
Incorrect Responders	4	6	4	4	6

Note: The number in each cell shows the number of questions on which the responders' FPRT was longer. TR = Text relevant; TIR = Text irrelevant; QS = Stem of the question; CO = Correct option; INCO = Incorrect options.

The first hypothesis predicted that the correct responders would fixate longer on the TR, QS and CO AoIs. This hypothesis seems to be supported to a certain extent since in seven of the questions the correct responders had longer fixations in these areas while in four of the questions the incorrect responders did. Similarly, it was expected that the correct responders would fixate shorter on the TIR and INCO AoIs. The data indicate that in five of the questions the correct responders had longer fixations in these areas whereas in six questions the incorrect responders did. To conclude, the analyses on the FPRT indicate that it is not possible to suggest that EM behaviors of the correct and incorrect responders show clear-cut differences.

5.2.2 The TFC

Descriptive statistics for the TFC for literal questions (see Appendix D) indicate that the correct responders had more frequent fixations in the TR AoI of Question 6, 19, and 3, the QS AoI of Question 6 and 19, and the CO AoI of Question 6. In addition, they had fewer fixations in the TIR AoI of all the questions except Question 6 and the INCO AoI of Question 34, 31, 19, 3, and 14.

As for the TFC for inferential questions (see Appendix E) in the TR AoI, the correct responders made fewer fixations on all of the questions than the incorrect responders, contrary to the expectation. However, they fixated more on the QS AoI of Question 13, 36 and 17. The TFC results related to the CO AoI showed that the correct responders had more fixations than the incorrect responders on all of the five questions. In the TIR AoI of Question 28, 33, and 36, the incorrect responders had more fixations than the correct responders; only in the TIR AoI of Question 13 and 17 they had fewer fixations compared to the correct responders. Also, it was observed that the incorrect responders had more fixations than the correct responders in all of the questions’ INCO AoI, except Question 33.

Table 2 displays a summary of the analyses related to the TFC for literal and inferential questions. It was hypothesized that the correct responders would fixate more frequently on the TR, QS and CO AoIs. However, the data show that in three questions the correct responders and in eight questions the incorrect responders had more frequent fixations in the TR AoI, as opposed to what was expected. A similar behavior was observed in terms of the QS as well since the incorrect responders had more frequent fixations in six questions while the correct responders in five questions. Only in terms of the CO did the correct responders have more frequent fixations in six questions with the incorrect responders having frequent fixations in five questions.

Table 2. The Summary Table for the TFC

	AoIs				
	TR	TIR	QS	CO	INCO
Literal Questions					
Correct Responders	3	1	2	1	1
Incorrect Responders	3	5	4	5	5
Inferential Questions					
Correct Responders	0	2	3	5	1
Incorrect Responders	5	3	2	0	4
Total					
Correct Responders	3	3	5	6	2
Incorrect Responders	8	8	6	5	9

Note: The number in each cell shows the number of questions on which the responders' TFC was more frequent. TR = Text relevant; TIR = Text irrelevant; QS = Stem of the question; CO = Correct option; INCO = Incorrect options.

As for the TIR and the INCO AoIs, the hypothesis was that the correct responders would fixate less frequently than the incorrect responders in these areas.

This hypothesis seems to be supported to a greater extent as in the TIR AoI of eight questions and the INCO AoI of nine questions the latter had more frequent fixations.

To conclude, the analyses of the TFC provide inconsistent support for the differences between the correct and incorrect responders. Thus, it is not possible to suggest that the TFC of the correct and incorrect responders show clear-cut differences.

5.2.3 The TRT

Descriptive statistics for the TRT for literal questions are displayed in Appendix F. In the TR AoI of Question 6, 19, and 3, the correct responders' TRT was longer than that of the incorrect responders. However, in the TIR AoI of all of the questions except Question 6, the incorrect responders' TRT was longer than that of the incorrect responders.

In the QS AoI of Question 6 and 19, the correct responders spent more time than the incorrect responders did, and the incorrect responders spent more time in the QS AoI of other four questions, 34, 31, 3, and 14.

Measures regarding the TRT in the CO and INCO AoIs showed that compared to the incorrect responders, the correct responders spent more time while reading the CO of Question 6 and 14, and less time while reading the INCO of Question 34, 31, 19, 3, and 14.

The descriptive statistics for the TRT related to the inferential questions are displayed in Appendix G. It was found that compared to the incorrect responders, the correct responders spent less time in the TR AoI of all five questions, namely Question 28, 33, 13, 36, and 17, contrary to the expectation. In addition, only in the TIR AoI of Question 28 and 36, the correct responders had shorter fixations than the incorrect responders. As for the QS AoI, the correct responders spent more time than the incorrect responders on Question 13, 36, and 17.

The TRT measures in the CO and INCO AoIs indicated that, compared to the incorrect responders, the correct responders spent more time in the CO AoI of Question 28, 33, 13, and 17 (all of the questions except Question 36), and less time in the INCO AoI of Question 28, 36, and 17.

Table 3 displays a summary of the analyses related to the TRT for literal and inferential questions. Contrary to the expectations, the incorrect responders' TRT was longer in eight of the questions in the TR AoI and in six of the questions in the QS AoI. The data regarding the TIR AoI, the INCO AoI, and the CO AoI are more in line with the expectations as the incorrect responders had longer reading time in the TIR AoI of seven questions and the INCO AoI of eight questions, and they had shorter reading time in the CO AoI of five questions.

Table 3. The Summary Table for the TRT

	AoIs				
	TR	TIR	QS	CO	INCO
Literal Questions					
Correct Responders	3	1	2	2	1
Incorrect Responders	3	5	4	4	5
Inferential Questions					
Correct Responders	0	3	3	4	2
Incorrect Responders	5	2	2	1	3
Total					
Correct Responders	3	4	5	6	3
Incorrect Responders	8	7	6	5	8

Note: The number in each cell shows the quantity of the questions on which the responders' TRT was longer. TR = Text relevant; TIR = Text irrelevant; QS = Stem of the question; CO = Correct option; INCO = Incorrect options.

To conclude, similar to the FPRT and the TFC, the data regarding the TRT do not provide consistent support to distinguish the EMs of the correct and incorrect responders.

5.3 Comparison of EM measures between literal and inferential questions

Table 4 provides the descriptive statistics for the FPRT in five AoIs for literal and inferential questions. And, it suggests that the percentage of FPRT in the QS AoI was the highest compared to the other regions for both the literal and inferential questions. In addition, it seems the participants spent more time on the part of the text relevant to the question in the literal questions compared to the inferential questions. The same is true for the part of the text irrelevant to the question as well.

Table 4. Descriptive Statistics for the Percentage FPRT across the AoIs by Type of Question

	Literal		Inferential	
AoI	M	SD	M	SD
Text relevant	13.64	4.85	9.08	3.19
Text irrelevant	10.11	2.69	4.40	1.59
Question stem	16.39	4.60	15.39	5.86
Correct option	7.44	2.26	8.57	2.41
Incorrect options	6.75	1.93	8.23	2.00

Note: N = 32.

A 2 x 5 repeated measures ANOVA with type of question (literal, inferential) and AoI (TR, TIR, QS, CO, INCO) as within group factors indicated a significant interaction between the two factors with a large effect size (see Table 5).

Table 5. ANOVA Summary Table for the Effects of Question Type and AoI on FPRT

Source	SS	Df	MS	F	Partial η^2
Type of question (TQ)	217.30	1	217.30	49.63***	.639
Error	122.59	28	4.38		
AoI	3155.74	4	788.94	36.12***	.563
Error	2446.29	112	21.84		
TQ * AoI	620.91	4	155.23	24.06***	.462
Error	722.64	112	6.45		

Note: *** p < .001.

The interaction between type of question and AoI can be seen in Figure 7, which shows clear-cut differences between the literal and inferential questions in terms of the FPRT in the TR and TIR AoIs. In both AoIs, the participants spent more time while they were engaged in answering the literal questions. As for, the FPRT in the QS AoI, the CO AoI, and the INCO AoI, significant differences were not observed between the two types of questions. It should be noted that for both types of questions, the longest FPRT was in the QS AoI compared to the other AoIs.

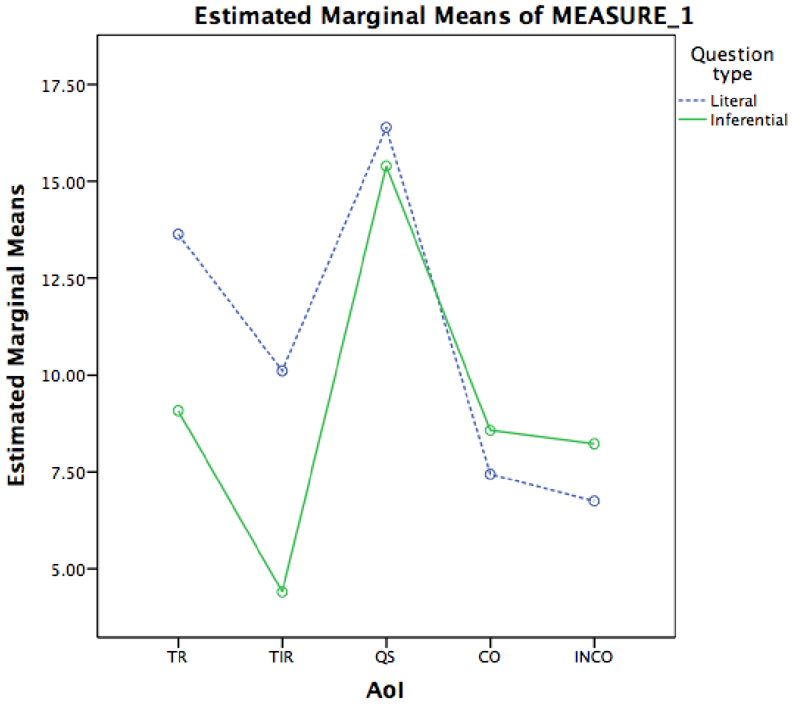


Fig. 7. Interaction between type of question and AoI in terms of FPRT.

Table 6 provides the descriptive statistics for the TFC across the AoIs for the literal and inferential questions. It seems the most frequent fixations were on the part of the text irrelevant to the question in the literal questions, whereas it was on the part of the text relevant to the question in the inferential questions.

Table 6. Descriptive Statistics for TFC across the AoIs by Type of Question

AoI	Literal		Inferential	
	M	SD	M	SD
Text relevant	18.15	5.19	14.41	5.45
Text irrelevant	24.62	7.34	6.32	3.69
Question stem	6.59	3.83	6.04	4.12
Correct option	3.25	1.38	4.39	1.78
Incorrect options	7.45	3.71	8.77	3.32

Note: N = 32.

A 2 x 5 repeated measures ANOVA with type of question (literal, inferential) and AoI (TR, TIR, QS, CO, INCO) as within group factors indicated a significant interaction between the two factors with a large effect size (see Table 7).

Table 7. ANOVA Summary Table for the Effects of Question Type and AoI on TFC

Source	SS	Df	MS	F	Partial η^2
Type of question (TQ)	1174.70	1	1174.70	68.02***	.708
Error	483.59	28	17.27		
AoI	7230.64	4	1807.66	53.79***	.658
Error	3763.98	112	33.61		
TQ * AoI	3934.39	4	983.59	115.98***	.806
Error	949.86	112	8.48		

Note: *** p < .001.

The interaction between type of question and AoI can be seen in Figure 8, which shows clear-cut differences between the literal and inferential questions in terms of the TFC in both the TR and TIR AoIs. Specifically, more frequent fixations were observed in both areas in the literal questions. It is worth noting that the TFC in the TIR AoI increased for the literal questions whereas it decreased for the inferential questions. As for the TFC in the QS AoI, the CO AoI, and the INCO AoI, significant differences were not observed between the two types of questions.

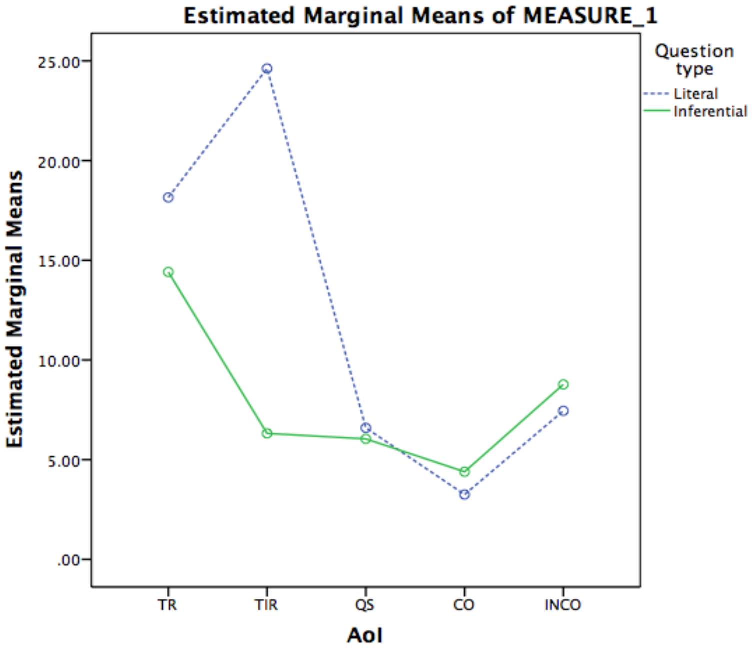


Fig. 8. Interaction between type of question and AoI in terms of TFC.

Table 8 provides the descriptive statistics for the TRT in five AoIs for the literal and inferential questions. The data indicated that the TRT for the literal questions was higher than that for the inferential questions in terms of both the TR and TIR AoIs.

Table 8. Descriptive Statistics for TRT across the AoIs by Type of Question

AoI	Literal		Inferential	
	M	SD	M	SD
Text relevant	17.83	6.20	13.55	5.74
Text irrelevant	25.68	12.55	5.53	3.32
Question stem	6.36	4.06	5.85	4.21
Correct option	3.44	1.71	4.59	1.96
Incorrect options	7.38	4.52	9.78	4.27

Note: N = 32.

A 2 x 5 repeated measures ANOVA with type of question (literal, inferential) and AoI (TR, TIR, QS, CO, INCO) as within group factors indicated a significant interaction between the two factors with a large effect size (see Table 9).

Table 9. ANOVA Summary Table for the Effects of Question Type and AoI on TRT

Source	SS	Df	MS	F	Partial η^2
Type of question (TQ)	1326.62	1	1326.62	42.07***	.600
Error	882.94	28	31.53		
AoI	6776.22	4	1694.06	32.79***	.539
Error	5786.45	112	51.66		
TQ * AoI	4933.87	4	1233.47	60.713***	.684
Error	2275.43	112	20.32		

Note: *** p < .001.

The interaction between type of question and AoI can be seen in Figure 9, which indicates that the TRT for the literal questions was higher in both the TR and TIR AoIs. It should be noted that the TRT in the TIR AoI increased for the literal questions whereas it decreased for the inferential questions. As for the TRT in the QS AoI, the CO AoI, and the INCO AoI, significant differences were not observed between the two types of questions.

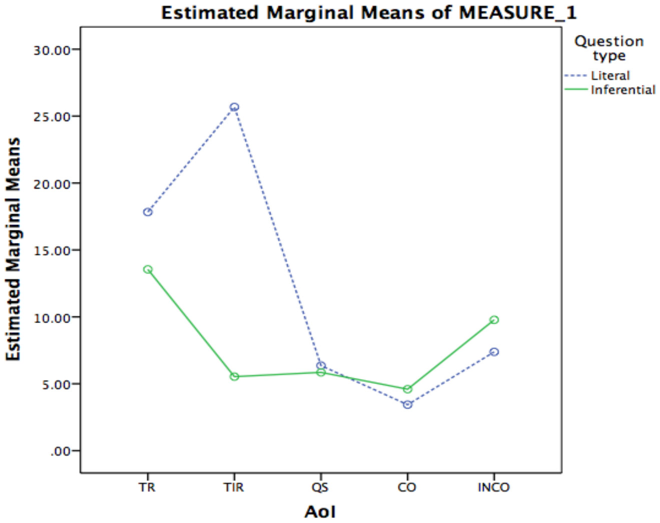


Fig. 9. Interaction between type of question and AoI in terms of TRT.

To summarize, the proportions of the FPRT and the TRT as well as the TFC indicate that the differences between the two types of questions are apparent in the TR and TIR AoIs, in line with Hypothesis 2. However, in terms of the QS, CO and INCO AoIs, differences between the two question types are not notable.

5.4 The first 10 fixations for the questions

In order to have further understanding about the participants' reading behaviors and strategy use, the order of the first 10 fixations for the literal and inferential questions were analyzed as it was thought that the sequence of fewer fixations would not be informative enough to have an idea about the strategy use in large AoIs.

The analysis of the first 10 fixations for the literal questions (see Appendix H) revealed that 79.7% of the correct responders and 81.4% of the incorrect responders read the QS before they read the text carefully. Thus, it can be said that a sizeable majority of the participants (80.2%) read the QS before reading the text. Moreover, the analysis also showed that 76.8% of the correct responders and 77% of the incorrect responders read the options before reading the passage. In other words, as a reading strategy, most of the participants (77.0%) read the options before reading the passage.

The summary analysis of the first 10 fixations for the inferential questions (see Appendix I) showed that 89.3% of the correct responders and 78.9% of the incorrect responders read the QS before they read the text carefully. Thus, it can be said that the majority of the participants (85.6%) preferred reading the QS before trying to find the TR AoI. As being another reading strategy, 85.6% of the participants read the options before reading the text. More specifically, 89.3% of the correct responders and 78.9% of the incorrect responders did not start to read the text without reading the options.

5.5 Relationships between reading performance and EM measures

Table 10 provides Pearson Product-Moment Correlations of overall reading comprehension scores (*M* = 30.34, *SD* = 2.89, *Min.* = 26, *Max.* = 36) with EM measures (FPRT, TFC, and TRT) in the five AoIs (TR, TIR, QS, CO, INCO). It was observed that the reading comprehension scores did not correlate with any of the EM measures for either literal or inferential questions.

Table 10. Correlations of Overall Reading Scores with EMs across the AoIs

	FPRT	TFC	TRT
Text relevant AoI for literal questions	.294	.075	.095
Text irrelevant AoI for literal questions	.097	-.097	-.047
Question stem AoI for literal questions	-.156	.040	-.030
Correct option AoI for literal questions	.034	.113	.106
Incorrect options AoI for literal questions	.131	-.006	-.069
Text relevant AoI for inferential questions	-.245	-.099	-.099
Text irrelevant AoI for inferential questions	.060	.235	.240
Question stem AoI for inferential questions	-.043	.087	.092
Correct option AoI for inferential questions	-.157	-.026	-.090
Incorrect options AoI for inferential questions	.008	-.181	-.045

Note: FPRT = First pass reading time, TFC = Total fixation count, TRT = Total reading time; r > .05 for all correlations; N = 32.

CHAPTER 6

DISCUSSION

6.1 Comparison of the EMs of correct and incorrect responders

In non-test conditions, readers pass through levels of comprehension, starting with lower-level skills such as phonological awareness, alphabetic understanding, and fluency (Dole et al., 1991), and connecting them with higher-level skills such as automaticity in reading and schema construction to comprehend the text (Basaraba et al., 2013). Accordingly, readers go through three levels of comprehension – literal, inferential and evaluative comprehension – in non-test conditions (Graesser et al., 1997; Kintsch, 1988, 1998; Lorch & van den Broek, 1997). Thus, they are expected to do so while taking a cognitively valid reading test (Bax & Weir, 2012; Weir, 2005). In view of the fact that cognitive validity of a test depends on the construal of the correct responses to questions, correct responses and processes triggering them are important factors for cognitive validity (Alderson, 2000). Thus, it is imperative to examine the cognitive processing gone through the completion of the task.

Eye-tracking is considered to provide an effective means to study text processing strategies as it is an on-line tool allowing collection of different indices of processing simultaneously (Hyönä & Lorch, 2004) which is not likely via other on-line or off-line methods. As such, in order to gain insights into cognitive processing of test takers while they are taking a reading test, eye-tracking methodology was used in the current study. Especially, the differences between test takers answering a question correctly and those incorrectly in terms of the FPRT, TFC and TRT in the TR, TIR, QS, CO, and INCO AoIs were investigated in detail while they were answering literal and inferential questions.

An item analysis on the participants' behavioral data indicated that the test was easy for the participants, so the items did not have discrimination power. Although there was a time constraint to complete the reading test, most of the participants answered the questions correctly. In other words, the NDRT items did not differentiate well between the students with higher and lower reading scores because they were too easy for the participants of the study. In addition the alpha coefficients that provide information about the internal consistency reliability of the test were quite below the desired level of .70 (Bachman, 2004) both for the test as a whole and for the literal and inferential questions separately.

Process data through EMs were collected in order to gain further insights into how the items functioned. Specifically, the focus was on whether the participants' EMs differed in different AoIs depending on whether they gave the right or the wrong answer. It was expected that the participants' EMs would show differences if the items discriminated between right or wrong scorers. Previous research has shown that good readers have fewer backward regressions leading to longer forward first pass fixations, as they do not have any comprehension difficulties (Blanchard & Iran-Nejad, 1987; Hyönä et al., 2002; Pang; 2008; Rayner, 1998; Rayner & Frazier, 1987; Rayner et al., 2012). In addition, Kaakinen et al. (2003) state that longer first pass forward fixations would be measured if the part that is being read is relevant to the reading perspective of the readers. In the light of these studies, it was expected that the correct responders would have longer FPRT in the TR AoI compared to the incorrect responders. The results of the study confirmed the expectation to a certain extent because in seven of the questions the correct responders had longer fixations in the TR AoI while in four of the questions the incorrect responders did. The first reason for longer FPRT would be in line with Kaakinen et al.'s (2003) argument which stated that longer FPRT in the TR AoI on the side of the correct responders would because of fewer comprehension difficulties while reading, resulting in fewer regressions. Additionally, the correct responders might have adjusted their reading speed by slowing it down when they thought that they had found the TR area which should be read slowly but carefully to make sure they understand what they are reading in order to answer the question correctly, and this could be the second reason why the correct responders had longer FPRT compared to the incorrect responders in the TR AoI.

According to Pang (2008), the correct responders are more strategic and more competent readers, so when they understand that the area of text being read is irrelevant to the question asked, they stop reading the part. Thus, compared to the FPRT of the incorrect responders, shorter FPRT was expected from the correct responders in the TIR AoI. The results regarding the FPRT in the TIR AoI verified the expectation partly because in five of the questions the correct responders had longer fixations in the TIR AoI while in six of the questions the incorrect responders did. On the other hand, supporting the Pang's (2008) argument, the incorrect responders might not quickly decide if the part being read was relevant or irrelevant, so it would be the reason why they continued to read the TIR area of six questions instead of ceasing reading.

It was assumed that in the QS AoI the correct responders' FPRT would be longer than the FPRT of the incorrect responders as it is known that more efficient readers are likely to have longer forward saccades (Rayner et al., 2012) as they do not face comprehension difficulties, and they slow down when they recognize the important nature of the information being read (Kaakinen et al., 2002). However, this expectation was also supported to some extent because in seven of the questions the correct responders had longer fixations in the QS AoI while in four of the questions the incorrect responders did. It could be said that the correct responders read the QS by slowing down, most probably, to give room for further cognitive processing, and they did longer forward saccades as they might not have faced as many comprehension difficulties as the incorrect responders did while reading. Also, because the participants were aware of the importance of reading the QS in order to answer the question correctly, the incorrect responders might have preferred to read the QS by slowing down to give themselves time for further processing similar to what the correct responders did, although it seems that this strategy awareness did not work for them to answer the question correctly.

For the same reason stated in the discussion of reasons for longer FPRT in the QS AoI, compared to the incorrect responders the correct responders were expected to have longer FPRT in the CO AoI but shorter FPRT in the INCO AoI. The expectation was supported to some extent as in seven of the questions the correct responders had longer fixations in the CO AoI while in four of the questions the incorrect responders did, and even minimizing the difference, in five of the questions the correct responders had longer fixations in the INCO AoI while in six of the questions the incorrect responders did. These results showed that both the correct and incorrect responders were aware of the importance of reading all of the options by paying close attention, which necessitates slower reading thus longer FPRT, to be able to answer the question correctly, although the correct strategy use was in vain for the incorrect responders.

All in all, the analyses on the FPRT showed that it is not possible to suggest that there is a clear-cut difference between the EM behaviors of the correct and incorrect responders.

The TFC was the second index that was investigated in specified five AoIs in order to see how many times the target area was fixated. Previous research has shown that multiple fixations and regressions may cause an increase on the TFC on certain areas of texts and indicate processing difficulties and comprehension failures (Rayner & Pollatsek, 1989; Roberts & Siyanova-Chanturia, 2013), or cognitive load associated with the part being processed (Bax, 2013; Bax & Weir, 2012; Blanchard & Iran-Nejad, 1987; Clifton et al., 2003; Rayner et al., 2006; Solheim & Uppstad, 2011). Thus, the correct responders were expected to make more fixations in the TR, QS, and CO AoIs but fewer fixations in the TIR and INCO AoIs compared to the incorrect responders. However, the expectation was not confirmed by the results of the current study. The correct responders fixated fewer than the incorrect responders did in the TR AoI. It seems that compared to the correct responders, the incorrect responders fixated more in the TR AoI because of the comprehension difficulties they faced, supporting the results of the first group of studies (Rayner & Pollatsek, 1989; Roberts & Siyanova-Chanturia, 2013) although this did not help them to answer the question correctly. Because of the difficulty faced, the incorrect responders might go back and forth in the text to find relationships among ideas, and reread to increase their understanding when the text or finding the related information to the question was difficult for them. It is clear that the use of these reading strategies increases the TFC in the target area. For that reason, it would be said that despite more fixations in the target area, the incorrect responders could not answer the question correctly because of the demanding cognitive processing. More specifically, unknown vocabulary included and/or the grammatical complexity of the part being read might have created difficulty in connecting the information proposed by the text by making the comprehension more difficult to them, and necessitating the incorrect responders make more forward and backward fixations in the TR AoI. Despite the occurrence of more fixations in the TR AoI, the lack of deep analysis and cognitive processing of the information in the area and/or the inability to activate background information to get help for better comprehension might have led them to answer the question incorrectly. In contrast, the fewer fixations by the correct responders in the TR AoI would be the signals of the lack of comprehension difficulty in the area, which showed that they did not face as many difficulties as the incorrect responders did while decoding the information included in the area.

As for the TIR AoI, the correct responders were expected to fixate less than the incorrect responders did, since they were expected to realize the irrelevant nature of the part being read to the question asked, and then, quit reading or skim (instead of careful reading) that part by being strategic readers (Pang, 2008), resulting in fewer fixations in the area. The results validated the expectation to a certain extent, as the correct responders’ TFC was fewer in eight questions. Because the correct responders were more strategic readers, immediately after identifying the irrelevant nature of the part, they might have quit reading not to spend any time in vain.

On the QS AoI, the correct responders were expected to fixate more than the incorrect responders did as a result of significance of the area (Bax, 2013; Bax & Weir, 2012; Blanchard & Iran-Nejad, 1987; Clifton et al., 2003; Rayner et al., 2006; Solheim & Uppstad, 2011). However, the expectation could not be approved. The incorrect responders had more frequent fixations in six questions while the correct responders had more frequent fixations in the other five questions. The similar behavior of the correct and incorrect responders would be as a result of a reading strategy. The participants had passed the YDS in English 3 to 8 months before becoming a regular ELT student at Boğaziçi University, meaning just a short time prior to becoming a participant in the current study. Thus, they were used to reading short passages and answering questions related to texts with the use of different reading strategies, also, they were knowledgeable about the importance of reading the QS carefully to be able to answer the question correctly. For that reason, both the correct and incorrect responders might go back and forth while reading the QS to make sure that they understood what is asked or between the TR/TIR AoI and the QS AoI to feel confident that they made correct connections between the text and the question, though the correct strategy use did not help the incorrect responders give the right answer. This was also confirmed through the examination of the first 10 fixations, which showed that both the correct and incorrect responders read the questions before referring to the text.

Considering arguments mentioned above (Bax, 2013; Bax & Weir, 2012; Blanchard & Iran-Nejad, 1987; Clifton et al., 2003; Kintsch, 1998; Rayner et al., 2006; Solheim & Uppstad, 2011), the correct responders were expected to fixate more than the incorrect responders did on the CO AoI. The results indicated that there was not a significant difference between the TFC of the correct and incorrect responders in the CO AoI, as the correct responders had more fixations in six questions while the incorrect responders had more fixations in five questions. It may probably be because of the fact that when the correct responders thought they had found the correct option, they did not choose it in an instant, but instead they gave themselves some time to think more, to go back to the TR/TIR AoI and then to reread the CO, leading to the increase in the TFC. On the other hand, the incorrect responders might have needed to go back and forth between the CO and TR/TIR AoIs many times to connect the ideas to each other in order to find the correct answer.

The last expectation related to the TFC was set in the light of Kintsch's (1998) and Pang's (2008) arguments, which posit that by being more strategic the correct responders stop focusing on the area when they realize that the part being read does not help them to answer the question correctly. Accordingly, the correct responders were expected to fixate fewer than the incorrect responders did in the INCO AoI. Supporting it, the results showed that the correct responders had fewer fixations than the incorrect responders in nine of the questions out of eleven. Thus, when the correct responders realized the options were distractors, they probably ceased focusing on the area. In contrast, the same EM behavior and strategy use could not be observed on the side of the incorrect responders, except in two questions. The incorrect responders must have been distracted by the INCO showing that the distractors worked properly for them strengthening the validity of the test. Thus, they must have felt the necessity to make many fixations in the INCO AoI, or regressions between the TR/TIR AoI and the INCO AoI to be able to fulfill the task requirement, which was answering the question correctly. In brief, it would be said that the reason for more TFC by the incorrect responders in the INCO AoI was due to difficulty in cognitive processing, which is in parallel with arguments of the researchers in the first group (Rayner & Pollatsek, 1989; Roberts & Siyanova-Chanturia, 2013).

To sum up, because the analyses on the TFC presented inconsistent support for the differences between the correct and incorrect responders, it was not possible to argue that there were clear-cut differences between the TFC of the correct and incorrect responders.

The TRT index includes all of the fixations observed on the target, so it divulges the total amount of time the participant spent in that AoI (Roberts & Siyanova-Chanturia, 2013; Siyanova-Chanturia et al., 2011a; Siyanova-Chanturia et al., 2011b). Taking into account Rayner’s (1998) argument which states that it is better to measure TRT as a major EM measure when the target AoI is larger than a single-word, the third index examined in identified five AoIs was the TRT in the current study as the TRT is measured based on all of the fixations in a given area including the FPRT and may indicate the time that the FPRT and the time which may have been spent to overcome processing difficulties (Liversedge, Paterson, & Pickering 1998).

Considering that the task relevance (Kaakinen et al. 2002, Solheim & Uppstad, 2011) and cognitively loaded nature of the target AoI (Bax, 2013; Blanchard & Iran-Nejad, 1987) have a boosting effect on the number and duration of fixations increasing the TRT, it was expected that the correct responders’ TRT would be longer than that of the incorrect responders in the TR AoI. However, the results showed that the correct responders’ TRT was less than the incorrect responders’ in eight questions, disproving the expectation. The first reason why the incorrect responders spent more time in the TR AoI could be because of comprehension difficulty they faced, supporting the results of Rayner & Pollatsek (1989) and Roberts and Siyanova-Chanturia (2013). The incorrect responders would feel the necessity to go back and forth in the text to find relationships among ideas, have time to think or guess the meaning of an unknown word, and reread to increase their understanding when the text or finding the related information to the question was difficult. On the contrary, the shorter TRT by the correct responders in the TR AoI could signal processing efficiency. In light of Kintsch’s (1998) theory of comprehension, it could be said that the correct responders might have focused on macropropositions to build a coherent mental model of the text as opposed to the incorrect responders who might have focused on minor details, which might have yielded longer fixation duration in the TR AoI.

The TIR AoI was the area that was irrelevant for successful performance on a given item. It was thought that the correct responders would spend less time than the incorrect responders in this area. The results of the study confirmed this expectation to a certain extent because the incorrect responders had longer reading time in the TIR AoI of seven questions, but shorter time in the TIR AoI of four questions. In general, it could be suggested that the correct responders might have realized the irrelevant nature of the area being read to the question, which could have made them stop reading or else to skim the area to use the time efficiently. It could also be that the correct responders did not have as much comprehension difficulty as the incorrect responders processing the TIR AoI.

Bax (2013), Bax & Weir (2012), Blanchard and Iran-Nejad (1987), Clifton et al. (2003), Rayner et al. (2006), and Solheim and Uppstad (2011) argue that cognitive load and the importance of the part being read augment the time spent on an area. In parallel with this notion, it was expected that the correct responders would spend more time than the incorrect responders in the QS AoI as they were assumed to be more strategic readers by being more alert on the prominence of the QS for successful task completion. Nevertheless, the results did not confirm the expectation. It was observed that although the correct responders spent more time in the QS AoI in five questions, the incorrect responders' TRT was longer in other six questions. The absence of the clear-cut difference between the correct and incorrect responders could be due to the use of test taking strategies. Because of their background, the participants were probably well-informed about the importance of reading the QS carefully for successful task completion.

The results of previous studies demonstrated that successful comprehenders spent more time to process important words than to process unimportant words (Kintsch, 1998; van der Schoot et al., 2008). In consideration of the results of these studies, the correct responders in the current study were expected to spend more time than the incorrect responders in the CO AoI. The results of the study partially confirmed the hypothesis, as the correct responders' TRT was longer in six questions but shorter than the incorrect responders' TRT in five other questions. It seems that as a reading strategy, the correct responders did not choose the correct answer immediately when they found it, but they gave themselves a thought for deeper processing and connecting each piece of information to each other in order to answer the question correctly. On the other hand, as a result of cognitive processing difficulty, the incorrect responders might feel the necessity to shuttle among the CO and TR/TIR AoIs many times to connect the ideas to each other in order to find the correct answer, and this could be the reason for having longer TRT in the CO AoI than the incorrect responders. In short, the strategy use by the correct responders and cognitive processing difficulty faced by the incorrect responders might be the reason for similar TRT in the CO AoI.

According to Kintsch's (1998) theory and Pang's (2008) study, successful readers are more strategic readers, so they continue reading if the area being read is relevant to the task, but if not, they stop focusing on the area being read not to spend time in vain (van der Schoot et al., 2008). Correspondingly, it was expected in the current study that the correct responders would spend less time than the incorrect responders in the INCO AoI, meaning their TRT would be shorter than the incorrect responders' TRT. Substantiating the expectation, the EM results indicated that the correct responders' TRT was shorter than that of the incorrect responders in the

INCO AoI in eight questions. The incorrect responders in the study must have been distracted by the INCO more than the correct responders since it seems that they had to spend more time in the INCO AoI, which was caused by frequent and longer fixations in the area, in order to answer the question correctly. In sum, it could be said that the incorrect responders had longer TRT than the correct responders' TRT in the INCO AoI on account of cognitive processing difficulty.

To conclude, similar to the FPRT and the TFC, the data related to the TRT did not give consistent support to differentiate the EMs of the correct and incorrect responders. As a result, supporting Solheim and Uppstad's (2011) argument, no clear-cut difference between the test takers answering the question correctly and those answering it incorrectly was observed. The fact that the EM data did not show consistent patterns could be due to the items having low discrimination power. In this sense, it can be concluded that the information obtained from the EM data matched that obtained from behavioral performance data.

6.2 Comparison of EM measures between literal and inferential questions

Another goal of the study was to investigate whether EMs showed differences depending on the type of question. Longer fixations and more frequent fixations were expected with the literal questions in all AoIs as they are considered to be text-bound (Kintsch, 1998). However, the results confirmed this expectation in terms of the TR and TIR AoIs, but not in terms of the QS, CO and INCO AoIs where no significant differences were observed.

With respect to all three measures of EM, the analyses indicated a significant interaction between the question type (literal vs. inferential) and AoI (TR, TIR, QS, CO, INCO) with a large effect size. It was seen that the percentage of the FPRT in the QS AoI was the highest compared to the other regions for both the literal and inferential questions. This could be because of the fact that the participants were aware of the importance of complete understanding of what was asked to find the correct answer of any kinds of questions. Additionally, clear-cut differences between the literal and inferential questions were observed in terms of the FPRT, TFC and TRT in the TR and TIR AoIs. More specifically a higher percentage of the FPRT in the TR AoI and the TIR AoI in the literal questions compared to the inferential questions was observed, and this could be related to the levels of comprehension. As mentioned in Chapter 2, readers should extract explicitly stated information in the text to be able to answer literal questions, which constitutes the first level of comprehension, literal comprehension (Carnine et al., 2010). For that reason they should focus more on the TR/TIR AoI in order to answer the literal questions. However, the same necessity is not applicable to the inferential questions, which stand at the second comprehension level requiring the readers interact with a text and use background information to make inferences about meaning not explicitly stated (Applegate et al., 2002). Thus, it can be said that the participants went through different levels of comprehension to find the correct answer to the literal and inferential questions, which is a factor to strengthen the cognitive validity of the NDRT.

Moreover, it was seen that the TFC and the TRT in the TIR AoI increased for the literal questions, but they decreased for the inferential questions. It seems that the participants preferred to read both the TR and TIR AoIs instead of focusing only on the TR AoI to make sure that they understood the passage and they had all the necessary information to answer the question correctly while engaging with the literal questions knowing that the answer of the question is explicitly stated in the text. However, because they might have used their background knowledge and general understanding of the text to answer the inferential questions, the same behavior could not be observed while they were trying to answer the inferential questions.

The first 10 fixations did not show any differences between the question types, as both the correct and incorrect responders first focused on the QS AoI before reading the text and the CO/INCO AoI before reading the text, confirming Hypothesis 3. It was observed that most of the participants read the question before reading the text (80.2% for literal questions, 85.6% for inferential questions), which supported Bax and Weir’s (2012) findings which indicated that 96.4% of the participants first read the question, then the text carefully. This behavior was observed, most probably, for the sake of finding the TR AoI as fast as possible, which is a reading strategy expected from competent readers. Competent readers are expected to set a purpose in mind while reading, so reading the QS first might be the best way to set a purpose and think about whether the content of the area being read fits the reading purpose, and the use of this strategy might help them to decide what to read closely and what to ignore for successful task completion in given time.

Moreover, it was traced that the majority of the participants first read the options carefully, and then they read the passages (77% for literal questions, 85.6% for inferential questions). This finding contradicts that of Nevo (1989) who showed that 95% of the participants read the passage carefully before reading the options. Instead of reading the passage first, the participants preferred to read the options most probably because of the limited time given for the task completion. The participants were to finish the task in given time, so they had to find a way to be quicker with the use of an appropriate reading strategy. Thus, it seems that both the correct and incorrect responders chose to have a look at the options first so that they would do search reading for the literal questions and could answer the questions without losing much time, or they would have an idea about what kind of information they were to find and where to focus on and read carefully to construct a schema in a short time for the inferential questions. As this reading behavior can be a characteristic of proficient readers, and imply higher level cognitive processing, it could be provided as evidence for the cognitive validity of the NDRT.

Additionally, the results of the study revealed that both the correct and incorrect responders read the entire passage thoroughly 90.9% of the time, substantiating Nevo's (1989) findings. Furthermore, all of the participants read more than a paragraph carefully, not bracing Bax and Weir's (2012) results precisely. These results demonstrated that almost all of the participants read both the TR and TIR AoIs of the texts carefully, and all of them read more than one paragraph instead of reading only the parts of the texts which was enough for them to answer the questions, showing that the participants were presumably checking whether their expeditious search reading had worked properly, so that they had not missed anything. Moreover, by being in line with the previous results (Bax & Weir, 2012), the results displayed that the incorrect responders read the CO carefully, which did not help them to answer the questions correctly though, and the correct responders read 63.6% of the CO by paying close attention. When analyzed in detail it was seen that, although they read the CO belonging to the inferential questions, they skimmed that of literal questions. It can be said that though the incorrect responders read the CO, they did not choose the CO because of comprehension difficulty, and the correct responders read the CO of inferential questions carefully as a result of cognitive load, but they did not deem it necessary to do the same thing for the literal questions' CO, as just skimming was enough for them to choose the CO as an answer.

To sum up, based on the first 10 fixations it was seen that the test takers answering a question correctly and those incorrectly do not differ in the AoIs they focus on while answering the literal and inferential questions. Both the correct and incorrect responders first focus on the QS AoI before reading the text and the CO/INCO AoI before reading the text, as expected. For that reason, Hypothesis 3 was verified by the study results.

6.3 Relationships between reading performance and EM measures

Previous research results showed that EMs could reflect reading performance (Bax, 2013; Bax & Weir, 2012). For that reason it was hypothesized that reading comprehension scores would correlate positively with EMs in the TR, QS, CO AoIs, but negatively with EMs in the TIR AoI and the INCO AoI. However, supporting Solheim and Uppstad's (2011) findings, the results uncovered that there was no correlation between the reading comprehension scores and any of the EM measures for the literal and inferential questions, disconfirming Hypothesis 4. The psychometric properties of the NDRT depending on the participants of the study could be the reason for the dissociation. The item analysis displayed that alpha coefficients were less than the desired level for a test to be accepted as reliable for the sample. Also, difficulty and discrimination analyses presented that the average item difficulty and item total correlations were not high enough, showing that the test did not differentiate well between the students with higher and lower reading scores because it was too easy for the participants of the study.

To conclude, these findings are partly consistent with the research hypotheses. Hypothesis 1 predicted that the participants answering a question correctly would have longer FPRT, and TRT with more TFC in the TR AoI, the QS AoI, and the CO AoI, but less FPRT, and TRT with fewer fixations in the TIR and INCO AoIs compared to those answering a question incorrectly. However, the analyses showed that it is not possible to suggest that there is a clear-cut difference between the EM behaviors of the correct and incorrect responders as a result of the use of the higher order reading strategies by both the correct and incorrect responders. Hypothesis 2 envisaged that the participants answering the literal questions would have longer FPRT and TRT with more frequent fixations in all five AoIs than the inferential questions. As a result of participants going through levels of comprehension and using variety of reading strategies, the second hypothesis was confirmed to some extent as the data results in the TR and TIR AoIs were in line with the hypothesis, but the hypothesis could not be verified with the results in the QS, CO and INCO AoIs. Hypothesis 3 anticipated that both the correct and incorrect responders would first focus on the QS AoI before reading the text, and the CO/INCO AoI before reading the text. It was seen that readers who are competent and experienced in reading tasks apply similar reading behaviors, thus they first focus on the QS AoI and the CO/INCO AoI before reading the text. Lastly, Hypothesis 4 predicted that reading comprehension scores would correlate positively with EMs in the TR, QS, and CO AoIs but negatively with EMs in the TIR and INCO AoIs. However, there could not be observed any correlation as a result of low level of psychometric properties of the NDRT. All in all, the results of the study validated some of the previous study results while negating some others.

CHAPTER 7

CONCLUSIONS

This study was first motivated to determine whether EM behaviors in different parts of the text and questions differed according to participants' performance (i.e., those who answered a question correctly vs. those who provided an incorrect answer) and the type of question (literal vs. inferential). The second aim of the study was to examine the relationship between reading comprehension and EM behaviors in different parts of the text and questions. Examining the results of the data, it was intended to have an idea about the cognitive processes gone through which are related to the cognitive validity of reading tasks.

Although eye-tracking has been used in educational research in recent years, it has been used mainly to analyze word processing, inferences, syntactic processing, and global discourse processing and these studies were mostly in first language (L1). Only a few of studies have investigated cognitive validity in L2. As there are only a handful of studies in educational research examining cognitive processing and validity in L2 through eye-tracking, this study was believed to clarify if the EM behaviors in different parts of the text and questions vary according to participants' performance and the type of question, and if reading comprehension correlates with EM behaviors in different parts of the text and questions. The reason for proposing eye-tracking methodology in the study was that it provided a better way to measure cognitive validity in that, unlike the available methods, it does not interfere with the natural test taking processes and gives richer information about the cognitive processes.

In order to find answers to the questions, psychometric properties of the NDRT in terms of the test’s internal consistency reliability, item difficulty and item analysis; differences between the participants answering a question correctly and those of incorrectly in terms of the FPRT, TFC, and TRT in the TR, TIR, QS, CO, and INCO AoIs; differences in the EMs (FPRT, TFC, and TRT) in five AoIs (TR, TIR, QS, CO, and INCO) depending on the question type (literal vs. inferential questions); the first 10 fixations; and the correlation between the reading comprehension scores and EMs (FPRT, TFC, and TRT) in the TR, TIR, QS, CO, and INCO AoIs were investigated in detail with the use of eye-tracking so that further insights would be attained into cognitive processing of test takers while they were taking an academic reading test.

These analyses first demonstrated that the NDRT does not discriminate sufficiently between the participants with higher and lower reading scores. Second, it was seen that when higher order reading strategies are used by both the correct and incorrect responders, the FPRT, TFC and TRT do not provide consistent support to distinguish the EMs of the correct and incorrect responders in the TR, TIR, QS, CO and INCO AoIs. In this regard, it can be concluded that behavioral outcome data and process-based EM data provide complementary results. Third, it was revealed that the EMs for the literal and inferential questions were rather different in terms of the TR and TIR AoIs, but they were not in the QS, CO and INCO AoIs, providing partial evidence for the cognitive validity of the NDRT. Fourth, the results proved that readers who are competent and experienced in reading tasks apply similar reading behaviors regardless of the type of question, thus they first focus on the QS AoI before reading the text and the CO/INCO AoI before reading the text. And finally, the low level of psychometric properties of the test resulted in lack of a relationship between the reading comprehension scores and EM measures for either the literal or inferential questions.

In the light of the evidence presented in the study it can be concluded that because the items necessitate the readers go through levels of comprehension while answering the literal and inferential questions, the set of items claim with some confidence to have cognitive validity in Khalifa and Weir’s (2009) terms. Additionally, it will be appropriate to argue that eye-tracking is a very effective on-line tool by compensating the possible limitations of traditional off-line data on participants cognitive processes and by giving way to collect different indices of processing simultaneously with high and spatial resolution to study cognitive processes of readers under test conditions and the ways in which test items can perform when eliciting particular cognitive processes in reading.

This study has several limitations. First, based on the psychometric properties of the NDRT, it can be said that the test was not entirely suitable for the sample. If a more suitable test had been used, EMs could have provided valuable insights regarding the cognitive processes of the participants while taking the test. Second, the participants were high proficiency learners of English with little performance differences. A study with a more heterogeneous group in terms of reading ability may provide different results. Third, EM data can be supported by another process-based data collection tool such as cued-retrospective reporting in order to triangulate the data. Finally, EMs can be investigated in relation to other factors such as gender, participants' content knowledge, and metacognitive strategy use to have a better understanding of how the test items function.

APPENDIX A

ITEM CHARACTERISTICS OF THE NDRT

	Frequency-Percent	Corrected Item-Total Correlation
Question 1	96.9	.024
Question 2	71.9	.142
Question 3	84.4	.077
Question 4	100.0	.000
Question 5	87.5	-.070
Question 6	53.1	.004
Question 7	90.6	.166
Question 8	84.4	.368
Question 9	96.9	.024
Question 10	100.0	.000
Question 11	9.4	.009
Question 12	71.9	.142
Question 13	75.0	.205
Question 14	84.4	-.045
Question 15	100.0	.000
Question 16	100.0	.000
Question 17	78.1	.080
Question 18	71.9	.329
Question 19	81.3	.092
Question 20	93.8	.129
Question 21	96.9	.152
Question 22	78.1	.220
Question 23	100.0	.000
Question 24	96.9	.152
Question 25	100.0	.000
Question 26	96.9	.216
Question 27	62.5	.206
Question 28	40.6	.019
Question 29	96.9	.152
Question 30	75.0	-.277
Question 31	65.6	.471
Question 32	81.3	-.023
Question 33	53.1	-.293
Question 34	62.5	.384
Question 35	87.5	.097
Question 36	75.0	.233
Question 37	71.9	.015
Question 38	62.5	.085

APPENDIX B

THE FPRT FOR LITERAL QUESTIONS

				Text Relevant		Text Irrelevant		Question Stem		Correct Option		Incorrect Option	
	% Correct	IT_CORR	N	M	SD	M	SD	M	SD	M	SD	M	SD
Question 6													
Correct responders	53	.004	17	140	.232	513	.590	931	.642	371	.149	263	.098
Incorrect responders			15	359	.447	394	.498	740	.475	313	.328	313	.216
Question 34													
Correct responders	63	.384	20	1030	1.881	547	.768	331	.183	301	.162	366	.343
Incorrect responders			12	973	1.263	810	.730	520	.379	318	.172	251	.159
Question 31													
Correct responders	66	.471	21	290	.565	118	.192	611	.580	413	.437	476	.379
Incorrect responders			11	377	.376	496	.566	851	.671	365	.253	353	.208
Question 19													
Correct responders	81	.092	26	1504	1.475	494	.668	764	.977	344	.327	324	.389
Incorrect responders			6	1095	1.274	458	.170	748	.655	203	.179	328	.148
Question 3													
Correct responders	84	.077	27	449	.654	258	.382	1361	.810	389	.322	329	.241
Incorrect responders			5	220	.301	397	.466	1067	.738	533	.719	404	.309
Question 14													
Correct responders	84	-.045	27	440	.478	521	.710	671	.493	357	.280	318	.203
Incorrect responders			5	460	.466	864	1.004	440	.174	240	.181	283	.139

Note: M values are in msec.; IT_CORR = Item total correlation.

APPENDIX C

THE FPRT FOR INFERENTIAL QUESTIONS

				Text Relevant		Text Irrelevant		Question Stem		Correct Option		Incorrect Option	
	% Correct	IT_CORR	N	M	SD	M	SD	M	SD	M	SD	M	SD
Question 28													
Correct responders	41	.019	13	415	.489	287	.613	433	.428	403	.379	380	.283
Incorrect responders			19	414	.316	274	.292	713	.501	373	.232	351	.311
Question 33													
Correct responders	53	-.293	17	208	.312	221	.296	605	.561	573	.515	368	.243
Incorrect responders			15	16	.258	144	.170	487	.357	269	.229	349	.236
Question 13													
Correct responders	75	.205	24	664	.626	255	.391	949	.740	279	.195	275	.176
Incorrect responders			8	648	.949	120	.194	740	.428	264	.269	448	.426
Question 36													
Correct responders	75	.233	24	529	.677	131	.189	686	.568	271	.205	179	.097
Incorrect responders			8	335	.573	560	.829	304	.214	308	.111	292	.169
Question 17													
Correct responders	78	.080	25	287	.353	561	.576	734	.552	408	.365	340	.436
Incorrect responders			7	703	.793	622	.556	953	.725	686	.431	505	.564

Note: M values are in msec.; IT_CORR = Item total correlation.

APPENDIX D

THE TFC FOR LITERAL QUESTIONS

				Text Relevant		Text Irrelevant		Question Stem		Correct Option		Incorrect Option	
	% Correct	IT_CORR	N	M	SD	M	SD	M	SD	M	SD	M	SD
Question 6													
Correct responders	53	.004	17	8.00	12.987	29.00	28.331	6.37	3.383	4.12	2.630	11.12	9.837
Incorrect responders			15	2.53	2.799	17.73	19.879	4.73	4.399	3.33	3.958	7.13	5.475
Question 34													
Correct responders	63	.384	20	43.89	40.417	27.36	25.882	6.78	9.840	4.00	3.480	7.78	6.241
Incorrect responders			12	62.83	36.620	37.83	33.310	8.16	4.239	8.75	7.136	13.50	10.247
Question 31													
Correct responders	66	.471	21	5.33	11.667	9.52	20.088	5.19	4.154	2.90	2.364	9.00	7.549
Incorrect responders			11	27.00	24.515	16.81	18.465	7.18	7.194	6.09	4.678	15.90	8.549
Question 19													
Correct responders	81	.092	26	55.19	42.752	13.34	20.746	8.57	7.365	2.38	2.593	11.84	9.379
Incorrect responders			6	25.66	30.539	35.00	54.310	8.50	9.049	3.33	3.141	15.50	15.162
Question 3													
Correct responders	84	.077	27	9.25	9.002	10.81	20.544	5.66	3.892	1.96	1.048	3.92	2.464
Incorrect responders			5	8.20	9.602	56.00	56.519	8.60	9.235	3.20	2.863	7.60	4.774
Question 14													
Correct responders	84	-.045	27	11.40	12.549	49.37	38.485	8.25	8.515	4.44	3.214	5.44	4.652
Incorrect responders			5	23.00	38.871	74.40	39.727	11.60	6.841	4.60	3.781	9.00	6.041

Note: M values are in msec.; IT_CORR = Item total correlation.

APPENDIX E

THE TFC FOR INFERENTIAL QUESTIONS

				Text Relevant		Text Irrelevant		Question Stem		Correct Option		Incorrect Option	
	% Correct	IT_CORR	N	M	SD	M	SD	M	SD	M	SD	M	SD
Question 28													
Correct responders	41	.019	13	7.61	8.057	5.69	7.728	5.69	6.675	6.00	4.102	11.69	9.384
Incorrect responders			19	17.52	17.401	7.36	10.724	8.47	9.287	3.15	2.630	12.63	7.896
Question 33													
Correct responders	53	-.293	17	6.94	14.506	8.94	19.311	4.41	3.873	6.00	4.541	11.05	9.051
Incorrect responders			15	12.57	21.713	10.00	19.775	6.50	6.560	3.42	3.955	10.14	7.440
Question 13													
Correct responders	75	.205	24	29.25	29.257	9.91	22.391	6.37	6.275	8.00	6.827	10.70	8.655
Incorrect responders			8	29.50	40.946	8.25	14.149	5.87	4.911	6.12	6.706	11.37	12.477
Question 36													
Correct responders	75	.233	24	21.62	18.761	5.70	11.961	4.12	2.938	3.91	3.091	4.66	3.806
Incorrect responders			8	24.12	28.073	8.37	7.981	3.87	3.563	3.87	2.748	7.25	6.881
Question 17													
Correct responders	78	.080	25	7.32	9.195	19.52	20.100	7.96	7.700	5.44	4.426	8.48	8.307
Incorrect responders			7	11.71	12.297	17.57	23.136	5.28	3.988	5.42	1.718	12.57	5.593

Note: M values are in msec.; IT_CORR = Item total correlation.

APPENDIX F

THE TRT FOR LITERAL QUESTIONS

				Text Relevant		Text Irrelevant		Question Stem		Correct Option		Incorrect Option	
	% Correct	IT_CORR	N	M	SD	M	SD	M	SD	M	SD	M	SD
Question 6													
Correct responders	53	.004	17	2549	3.816	9731	9.747	2651	1.774	1479	1.051	4854	5.293
Incorrect responders			15	1105	1.699	5746	6.576	1581	1.628	1146	1.160	2770	2.576
Question 34													
Correct responders	63	.384	20	13506	13.155	8719	8.643	2027	3.491	1218	1.106	2348	2.227
Incorrect responders			12	21473	13.128	13596	15.032	2822	1.561	2944	2.661	4610	3.357
Question 31													
Correct responders	66	.471	21	1341	2.986	2672	5.708	1522	1.268	1050	.989	3409	3.142
Incorrect responders			11	8002	6.915	4739	5.176	2149	2.081	2009	1.490	5447	3.096
Question 19													
Correct responders	81	.092	26	19042	16.448	4809	7.549	3169	2.623	878	.939	3908	3.621
Incorrect responders			6	10037	12.628	14225	25.261	2780	3.080	1273	1.593	5580	6.198
Question 3													
Correct responders	84	.077	27	3272	4.422	3469	6.702	2396	2.230	861	.669	1254	.871
Incorrect responders			5	2692	3.273	17931	15.049	3349	2.776	2158	2.524	4334	4.785
Question 14													
Correct responders	84	-.045	27	3415	4.183	15908	13.500	2647	3.075	1409	1.069	1610	1.641
Incorrect responders			5	7237	12.787	23663	14.561	3980	2.353	1391	1.221	3543	2.958

Note: M values are in msec.; IT_CORR = Item total correlation.

APPENDIX G

THE TRT FOR INFERENTIAL QUESTIONS

	% Correct	IT_CORR	N	Text Relevant		Text Irrelevant		Question Stem		Correct Option		Incorrect Option	
				M	SD	M	SD	M	SD	M	SD	M	SD
Question 28													
Correct responders	41	.019	13	2252	2.594	1687	2.351	1839	2.595	1975	1.362	4069	3.263
Incorrect responders			19	5253	5.280	1812	2.694	2787	3.156	1086	1.029	4629	3.312
Question 33													
Correct responders	53	-.293	17	2073	4.326	2563	5.324	1666	1.778	2886	2.459	5369	4.586
Incorrect responders			15	3710	6.414	2501	4.963	2359	3.151	1169	1.348	5047	3.589
Question 13													
Correct responders	75	.205	24	8831	9.524	2679	6.045	2191	2.593	2771	2.600	3509	3.168
Incorrect responders			8	9753	12.628	2047	3.721	1796	1.894	2414	2.514	3445	3.271
Question 36													
Correct responders	75	.233	24	6752	6.357	1780	3.751	1662	1.623	1051	.966	1232	1.143
Incorrect responders			8	6811	7.037	2788	2.920	1105	1.154	1067	.868	2337	2.436
Question 17													
Correct responders	78	.080	25	2158	2.624	5488	5.642	2480	2.770	1829	1.515	3068	3.219
Incorrect responders			7	3625	3.863	5241	6.122	1451	1.236	1780	.922	4871	2.360

Note: M values are in msec.; IT_CORR = Item total correlation.

APPENDIX H

THE SUMMARY ANALYSIS OF THE FIRST 10 FIXATIONS

FOR LITERAL QUESTIONS

		Did the participants read the question before reading the text? (at least three fixations)	Did the participants read the options before reading the text?
	N	N	N
Question 6			
Correct responders	17	15	13
Incorrect responders	15	12	10
Question 34			
Correct responders	20	14	12
Incorrect responders	12	10	9
Question 31			
Correct responders	21	18	19
Incorrect responders	11	8	9
Question 19			
Correct responders	26	19	20
Incorrect responders	6	4	4
Question 3			
Correct responders	27	24	22
Incorrect responders	5	5	5
Question 14			
Correct responders	27	20	20
Incorrect responders	5	5	5
		%	%
Correct responders		79.7	76.8
Incorrect responders		81.4	77.7
Total		80.2	77.0

Note: The first N column shows the number of correct and incorrect responders for each question; the second N column illustrates the number of correct and incorrect responders reading the question before the text for each question; the third N column displays the number of correct and incorrect responders reading the options before the text for each question.

APPENDIX I

THE SUMMARY ANALYSIS OF THE FIRST 10 FIXATIONS FOR INFERENTIAL QUESTIONS

		Did the participants read the question before reading the text? (at least three fixations)	Did the participants read the options before reading the text?
	N	N	N
Question 28			
Correct responders	13	10	11
Incorrect responders	19	14	13
Question 33			
Correct responders	17	17	15
Incorrect responders	15	13	12
Question 13			
Correct responders	24	22	22
Incorrect responders	8	7	7
Question 36			
Correct responders	24	21	23
Incorrect responders	8	5	7
Question 17			
Correct responders	25	22	21
Incorrect responders	7	6	6
		%	%
Correct responders		89.3	89.3
Incorrect responders		78.9	78.9
Total		85.6	85.6

Note: The first N column shows the number of correct and incorrect responders for each question; the second N column illustrates the number of correct and incorrect responders reading the question before the text for each question; the third N column displays the number of correct and incorrect responders reading the options before the text for each question.

REFERENCES

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

Amadieu, F., van Gogh, T., Paas, F., Tricot, A., & Mariné, C. (2009). Effects of prior knowledge and concept-map structure on disorientation, cognitive load, and learning. *Learning and Instruction, 19,* 376-386.

Andrews, S., Miller, B., & Rayner, K. (2004). Eye movements and morphological segmentation of compound words: There is a mouse in mousetrap. *European Journal of Cognitive Psychology, 16,* 285-311.

Applegate, M. D., Quinn, K. B., & Applegate, A. J. (2002). Levels of thinking required by comprehension questions in informal inventories. *The Reading Teacher, 56*, 174-180.

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

Balota, D. A., Flores d'Arcais, G. B., & Rayner, K. (Eds.). (1990). *Comprehension processes in reading.* Hillsdale, NJ: Erlbaum.

Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology, 17,* 364-390.

Basaraba, D., Yovanoff, P., Alonzo, J., & Tindal, G. (2013). Examining the structure of reading comprehension: Do literal, inferential, and evaluative comprehension truly exist? *Reading and Writing, 26*(3), 349-379.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing, 30*(4), 441-465.

Bax, S., & Weir, C. (2012). Investigating learners' cognitive processes during a computer-based CAE Reading test. *Cambridge ESOL: Research Notes, 47,* 3-14.

Besner, D., & Humphreys, G. W. (Eds.). (1991). *Basic processes in reading: Visual word recognition.* Hillsdale, NJ: Erlbaum.

Blanchard, H. E., & Iran-Nejad, A. (1987). Comprehension processes and eye movement patterns in the reading of surprise-ending stories. *Discourse Processes, 10,* 127-138.

Braze, D., Shankweiler, D., Ni, W., & Palumbo, L. C. (2002). Readers' eye movements distinguish anomalies of form and content. *Journal of Psycholinguistic Research, 31,* 25–44.

Brown, J. I., Fishco, V. H., & Hanna, G. (1993). *Nelson-Denny Reading Test Form G.* Itasca, IL: The Riverside Publishing Company.

Burton, C., & Daneman, M. (2007). Compensating for a limited working memory capacity during reading: Evidence from eye movements. *Reading Psychology, 28,* 163-186.

Buswell, G. T. (1935). *How people look at pictures.* Chicago: University of Chicago Press.

Carnine, D. W., Silbert, J., Kame'enuni, E. J., & Tarver, S. G. (2010). *Direct instruction reading* (5th ed.). Boston, MA: Merrill.

Carpenter, P. A., & Just, M. A. (1989). The role of working memory in language comprehension. In D. K. Klahr and K. Kotovsky (Eds.), *Complex information processing : The impact of Herbert A. Simon* (pp. 31-68). Hillsdale, NJ: L. Erlbaum Associates.

Carver, R. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading, 36*(2), 84-95.

Chaffin, R., Morris, R. K., & Seely, R. E. (2001). Learning new word meanings from context: A study of eye movements. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27,* 225-235.

Clifton, C., Staub, A. Jr., & Rayner, K. (2006). Eye movements in reading words and sentences. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 341-371). Oxford: Elsevier.

Clifton, C., Traxler, M. J., Mohamed, M. T., Williams, R. S., Morris, R. K., & Rayner, K. (2003). The use of thematic role information in parsing: Syntactic processing autonomy revisited. *Journal of Memory and Language, 49,* 317-334.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19,* 450-466.

Davey, B. (1988). Factors affecting the difficulty of reading comprehension items for successful and unsuccessful readers. *Journal of Experimental Education, 56,* 67-75.

Deutsch A., & Bentin, S. (2001). Syntactic and semantic factors in processing gender agreement in Hebrew: Evidence from ERPs and eye movements. *Journal of Memory and Language, 45*(2), 200-224.

de Greef, T., Botzer, A., & van Maanen, P. P. (2010). Eye-tracking = Reading the mind, in *Proceedings of the 28th Annual European Conference*, on Cognitive Ergonomics (ECCE 2010). Delft, Netherlands, August 25-27 2010, New York: ACM Press, 303-304.

Dillon, R. F. (1997). A new era in testing. In R. F. Dillon (Ed.), *Handbook on testing* (pp. 1-19). Westport, CN: Greenwood Press.

Dole, J. A., Duffy, G. G., Roehler, L. R., & Pearson, P. D. (1991). Moving from the old to the new: Research on reading comprehension instruction. *Review of Educational Research, 61*, 239-264.

Duchowski, A. T. (2003). *Eye tracking methodology: Theory and practice*. London: Springer.

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, *27*, 429-446.

Durkin, D. (1989). *Teaching them to read* (5th ed.). Boston, MA: Allyn & Bacon.

Dussias, P. E. (2010). Uses of eye-tracking data in second language sentence processing research. *Annual Review of Applied Linguistics, 30,* 149-166.

Ehrlich, S. F., & Rayner, K. (1983). Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning and Verbal Behavior*, *22*, 75-87.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), Examining listening: Research and practice in assessing second language listening, *Studies in Language Testing*, *35*, 77-151. Cambridge: Cambridge University Press.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology, 14,* 178-210.

Garrod, S., Freudenthal, S., & Boyle, E. (1994). The role of different types of anaphor in the on-line resolution of sentences in a discourse. *Journal of Memory and Language, 33,* 39-68.

Glenberg, M. A., Kruley, P., & Langston, W. E. (1994). Analogical processes in comprehension: Simulation of a mental model. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 600-609). San Diego, CA: Academic Press.

Goltz, T. T. (1975). *Comparision of the eye movements of skilled and less skilled readers* (Unpublished doctoral dissertation). Washington University.

Goodman, K. S. (1996). *On reading.* Portsmouth, NH: Heinemann.

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology, 48,* 163-189.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371-395.

Griffin, D. C., Walton, H. N., & Ives, V. (1974). Saccades as related to reading disorders. *Journal of Learning Disabilities, 7*, 310-316.

Guthrie, J. T., & Kirsch, I. S. (1987). Distinctions between reading comprehension and locating information in text. *Journal of Educational Psychology, 79*(3), 220-227.

Hannus, M., & Hyönä, J. (1999). Utilization of illustrations during learning of science textbook passages among low- and high-ability children. *Contemporary Educational Psychology, 24*, 95-123.

Hayes, S. C., White, D., & Bissett, R. T. (1998). Protocol analysis and the "silent dog" method of analyzing the impact of self-generated rules. *Analysis of Verbal Behavior, 15*, 57-63.

Heiman, J. R., & Ross, A. O. (1974). Saccadic eye movements and reading difficulties. *Journal of Abnormal Child Psychology, 2,* 53-61.

Herber, H. L. (1970). *Teaching reading in the content areas*. Englewood Cliffs, NJ: Prentice Hall.

Huey, E. B. (1908). *The psychology and pedagogy of reading.* New York: Macmillan.

Hyönä, J. (2010). The use of eye movements in the study of multimedia learning. *Learning and Instruction, 20*(2), 172-176.

Hyönä, J., & Lorch, R. F. (2004). Effects of topic headings on text processing: Evidence from adult readers' eye fixation patterns. *Learning and Instruction, 14*, 131-152.

Hyönä, J., Lorch, R. F., & Kaakinen, J. K. (2002). Individual differences in reading to summarize expository text: Evidence from eye fixation patterns. *Journal of Educational Psychology, 94*, 44-55.

Hyönä, J., & Pollatsek, A. (1998). Reading Finnish compound words: Eye fixations are affected by component morphemes. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 1612-1627.

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception and Psychophysics, 40*, 431-439.

Irwin, D. E. (1998). Lexical processing during saccadic eye movements. *Cognitive Psychology, 36,* 1-27.

Irwin, D. E., & Carlson-Radvansky, L. A. (1996). Cognitive suppression during saccadic eye movements. *Psychological Science, 7,* 83-88.

Juhasz, B. J., & Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29,* 1312-1318.

Juhasz, B. J., Starr, M. S., Inhoff, A. W., & Placke, L. (2003). The effects of morphology on the processing of compound words: Evidence from naming, lexical decisions and eye fixations. *British Journal of Psychology, 94*, 223-244.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87,* 329-354.

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension.* Boston: Allyn and Bacon.

Kaakinen, J. K., & Hyönä, J. (2007). Strategy use in the reading span test: An analysis of eye movements and reported encoding strategies. *Memory, 15*(6), 634-646.

Kaakinen, J. K., Hyönä, J., & Keenan, J. M. (2002). Perspective effects on online text processing. *Discourse Processes, 33,* 159-173.

Kaakinen, J. K., Hyönä, J., & Keenan, J. M. (2003). How prior knowledge, WMC, and relevance of information affect eye fixations in expository text. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 447-457.

Khalifa, H., & Weir, C. (2009). Examining reading: Research and practice in assessing second language reading. *Studies in Language Testing* (Vol. 29). Cambridge: UCLES.

King, A. (2007). Beyond literal comprehension: A strategy to promote deep understanding of text. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 267-290). Hillsdale, NJ: Lawrence Erlbaum.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A constructive-integration model. *Psychological Review, 95,* 163-182.

Kintsch, W. (1994). The psychology of discourse processing. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 721-740). San Diego, CA: Academic Press.

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

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

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of discourse comprehension and production. *Psychological Review, 85*, 363-394.

Kusela, H., & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *American Journal of Psychology, 113*(3), 387-404.

Leu, D. J., & Kinzer, C. K. (1999). *Effective literacy instruction (K-8)* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Liversedge, S., Paterson, K., & Pickering, M. (1998). Eye movements and measures of reading time. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 55-75). Oxford: Elsevier.

Lorch, R. F. Jr., & van den Broek, P. (1997). Understanding reading comprehension: Current and future contributions of cognitive science. *Contemporary Educational Psychology, 22*, 213-246.

Marslen-Wilson, W. (Ed.). (1989). *Lexical representation and process.* Cambridge, MA: MIT Press.

McCormick, S. (1992). Disabled readers' erroneous responses to inferential comprehension questions: Description and analyses. *Reading Research Quarterly, 27*(1), 54-77.

McKay, M. T., Fischler, I., & Dunn, B. R. (2003). Cognitive style and recall of text: An EEG analysis. *Learning and Individual Differences, 14*(1), 1-21.

Messick, S. (1976). Personality consistencies in cognition and creativity. In S. Messick (Ed.), *Individuality in learning* (pp. 4 -23). San Francisco: Jossey-Bass.

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

Murray, W. S., & Rowan, M. (1998). Early, mandatory, pragmatic processing. *Journal of Psycholinguistic Research, 27*(1), 1-22.

Nation, K. (2005). Children's reading comprehension difficulties. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 248-265). Oxford, MA: Blackwell.

National Institute of Child Health and Human Development. (2000). *Teaching children to read: An evidence-based assessment of scientific research literature on reading and its implications for instruction: Reports of the subgroups.* (NIH publication No. 00-4754). Washington, DC: U. S. Government Printing Office.

National Research Council. (1998). Organizational strategies for kindergarten and the primary grades. In C. E. Snow, M. S. Burns, & P. Griffin (Eds.), *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing, 6,* 199-215.

Ni, W., Fodor, J. D., Crain, S., & Shankweiler, D. (1998). Anomaly detection: Eye movement patterns. *Journal of Psycholinguistic Research, 27,* 515-539.

O’Brien, E. J., Raney, G. E., Albrecht, J. E., & Rayner, K. (1997). Processes involved in the resolution of explicit anaphors. *Discourse Processes, 23,* 1-24.

O’Brien, E. J., Shank, D. M., Myers, J. L., & Rayner, K. (1988). Elaborative inferences during reading: Do they occur on-line? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 410-420.

O’Brien de Ramirez, K. (2008). *Silent, oral, L1, L2, French and English reading through eye movements and miscues* (Unpublished doctoral dissertation). The University of Arizona, Tucson.

O'Regan, J. K. (1979). Moment to moment control of eye saccades as a function of textual parameters in reading. In P. A. Kolers, M. E. Wrolstad, & H. Bouma (Eds.), *Processing of visible language* (Vol. 1) (pp. 49-60). New York: Plenum.

Pang, J. (2008). Research on good and poor reader characteristics: Implications for L2 reading research in China. *Reading in a Foreign Language, 20*(1), 1-18.

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language, 41*, 427-456.

Pellgrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Perfetti, C. A. (1999). Comprehending written language: A blueprint of the reader. In C. M. Brown & P. Hagoart (Eds.), *The neurocognition of language* (pp. 167-208). New York: Oxford University Press.

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

Pugh, A. K. (1978). *Silent reading.* London: Heinemann.

Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin, 85,* 618-660.

Rayner, K. (1997). Eye movements, perceptual span, and reading disability. *Annals of Dyslexia, 47,* 30-52.

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

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology, 62*, 1457-1506.

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.

Rayner, K., & Duffy, S. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory and Cognition, 14*, 191-201.

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest, 2*, 31-74.

Rayner, K., & Frazier, L. (1987). Parsing temporarily ambiguous complements. *Quarterly Journal of Experimental Psychology, 39*, 657-673.

Rayner, K., & Frazier, L. (1989). Selection mechanisms in reading lexically ambiguous words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15,* 779-790.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading*. New Jersey: Prentice Hall.

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

Rayner, K., Sereno, S. C., & Raney, G. E. (1996). Eye movement control in reading: A comparison of two types of models. *Journal of Experimental Psychology Human Perception and Performance, 22*(5), 1188-1200.

Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*, 1290-1301.

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review, 3*(4), 504-509.

Roberts, L., & Siyanova-Chanturia, A. (2013). Using eye-tracking to investigate topics in L2 acquisition and L2 processing. *Studies in Second Language Acquisition, 35*, 213-235.

Rosenshine, B. V. (1980). Skills hierarchies in reading comprehension. In R. J. Spiro, B. C. Bruce, W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 535-559). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rubino, C. A., & Minden, H. A. (1973). An analysis of eye-movements in children with a reading disability. *Cortex, 9*, 217-220.

Rupley, W. H., & Blair, T. R. (1983). *Reading diagnosis and remediation: Classroom and clinic* (2nd ed.). Boston: Houghton Mifflin.

Salvucci, D., & Goldberg, J. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the Eye-tracking Research and Applications Symposium*, Palm Beach Gardens, FL, USA, 6-8 November 2000, New York: ACM Press, 71-78.

Sereno, S. C., O'Donnell, P. J., & Rayner, K. (2006). Eye movements and lexical ambiguity resolution: Investigating the subordinate bias effect. *Journal of Experimental Psychology: Human Perception and Performance, 32*, 335-350.

Singer, M. (1994). Discourse inference processes. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 479-515). San Diego, CA: Academic Press.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011a). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and nonnative speakers. *Second Language Research, 27*, 251-272.

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. (2011b). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multi-word sequences. *Journal of Experimental Psychology: Language, Memory, and Cognition, 37*, 776-784.

Snider, V. E. (1988). The role of prior knowledge in reading comprehension: A test with LD adolescents. *Direct Instruction News, 3*, 6-11.

Snow, C. E. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND.

Snow, C. E. (2003). Assessment of reading comprehension. In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 192-218). New York: Guilford.

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.

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.

Tai, R. H., Loehr, J. F., & Brigham, F. J. (2006). An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments. *International Journal of Research & Method in Education*, *29*(2), 185-208.

Taylor, S. E. (1962). An evaluation of forty-one trainees who had recently completed the "reading dynamics" program. *Eleventh Yearbook of the National Reading Conference*, 41-55.

Tinker, M. A. (1939). Reliability and validity of eye-movement measures of reading. *Journal of Experimental Psychology, 19,* 732-746.

Tinker, M. A. (1946). The study of eye movements in reading. *Psychological Bulletin, 43,* 93-120.

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

Vacca, J. L., Vacca, R. T., Gove, M. K., Burkey, L. C., Lenhart, L. A., & McKeon, C. A. (2009). *Reading and learning to read* (7th ed.). Boston, MA: Pearson.

van den Broek, P., Lorch, R. P., Linderholm, T. & Gustafson, M. (2001). The effects of readers' goals on inference generation and memory for texts. *Memory & Cognition, 29,* 1081-1087.

van der Schoot, M., Vasbinder, A. L., Horsley, T., & van Lieshout, E. C. D. M. (2008). The role of two reading strategies in text comprehension: An eye fixation study in primary school children. *Journal of Research in Reading*, *31*, 203-223.

van Dijk, T. A. (1980). *Macrostructures*. Hillsdale, NJ: Erlbaum.

van Hell, J. G., & Tokowicz, N. (2010). Event-related brain potentials and second language learning: Syntactic processing in late L2 learners at different L2 proficiency levels. *Second Language Research, 26,* 43-74.

Walker, L., Munro, J., & Rickards, F. W. (1998). Teaching inferential reading strategies through pictures. *Volta Review, 100,* 87-104.

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

Weir, C., Hawkey, R., Green, T., Devi, S. (2009). The cognitive processes underlying the academic reading construct as measured by IELTS. *British Council/IDP Australia IELTS Research Reports, 9*(4), 157-189.

Williams, R. S., & Morris, R. K. (2004). Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology, 16*, 312-339.

Winke, P., Gass, S., & Sydorenko, T. (2013). Factors influencing the use of captions by foreign language learners: An eye-tracking study. *The Modern Language Journal*, *97*(1), 254-275.

Zola, D. (1984). Redundancy and word perception during reading. *Perception & Psychophysics*, *36*, 277-284.

Zwaan, R. A., & Brown, C. M. (1996). The influence of language proficiency and comprehension skill on situation-model construction. *Discourse Processes, 21*, 289-327.