12-17-2020

Investigating Reading Behavior and Inference-making in Advanced L2 Reading Comprehension Assessment Tasks

Rurik Tywoniw

INVESTIGATING READING BEHAVIOR AND INFERENCE-MAKING IN ADVANCED L2 READING COMPREHENSION ASSESSMENT TASKS

by

RURIK TYWONIW

Under the Direction of Scott Crossley, PhD

ABSTRACT

Despite the ubiquity of reading comprehension tasks in English language proficiency tests (or sections of tests), the constructs underlying successful reading comprehension in English as a second/additional language at the advanced academic level are still not completely understood. Part of the reason for this gap in the current state of knowledge comes from how existing models of second language reading neglect higher-order reading skills. Many reading assessments overly target language proficiency skills and assume the transfer of first language literacy skills, leaving unexamined the higher-order skills of language learners who become skilled academic readers in their second or additional language. This study seeks to address the dearth of research on higher-order reading skills in advanced second language reading

comprehension by examining the activation of these skills in realistic L2 reading comprehension tasks. A reading comprehension test with three different tasks (MC questions, cloze, and summary) was developed and administered to 102 second language English and multilingual undergraduate and graduate students studying at a university in the US. Eye-movement behavior was recorded during these tasks, and each reading task was followed by a sentence verification task to measure activation of inferencing. Eye-movement behavior and inferencing are compared across the reading tasks, and additionally compared to language proficiency and reading comprehension scores. The tasks each elicited distinct patterns of reading behavior: the cloze task elicited careful local reading, the MC task elicited expeditious linear reading, and the summary task elicited both careful global reading and expeditious strategies. Cloze scores were closely related to language proficiency, but also related to reasoning ability and processing efficiency. MC scores were unrelated to proficiency. They were instead related more to reasoning ability and were predicted by readers' ability to efficiently process the MC questions. Inferencing ability was only predictive of score in the summary task. Summary scores were additionally influenced by global attention to the text, processing efficiency, reading motivation, and language proficiency. Implications for the use of each task as L2 reading assessment are discussed, as well as implications for the teaching of second language reading.

INDEX WORDS: Language Assessment, Reading Comprehension, Eye-tracking, Inferencing, Second Language Reading, Testing Reading

INVESTIGATING READING BEHAVIOR AND INFERENCE-MAKING IN ADVANCED L2 READING COMPREHENSION ASSESSMENT TASKS

by

RURIK TYWONIW

A Dissertation Submitted in Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2020

INVESTIGATING READING BEHAVIOR AND INFERENCE-MAKING IN ADVANCED L2 READING COMPREHENSION ASSESSMENT TASKS

by

RURIK TYWONIW

Committee Chair: Scott Crossley

Committee: Sara Cushing

Diane Belcher

Sarah Carlson

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

December 2020

DEDICATION

To everyone I consider family,

… if they'll have me.

ACKNOWLEDGMENTS

This dissertation has been a long time in the making, and although at times it seemed like it would never come together, with the help of many people this dissertation was able to find its feet and come together. A comprehensive expression of thanks to every person who contributed to this dissertation would be impossible, but I will attempt to spread the gratitude.

First and foremost, I would like to show my utmost thanks and appreciation to my advisor Scott Crossley. He has provided unquantifiable guidance, feedback, and support, giving both structure and latitude for exploring ideas. Without his presence as an advisor, I cannot imagine being able to pursue a topic which I am truly passionate about with the same level of insight and methodological support. I thank Scott for his patience, understanding, and mentorship, especially as I completed this dissertation during such a chaotic time.

I would also like to thank the members of my committee. I thank Sara Cushing, who has been my moral compass for investigating issues of language testing and has consistently been a font of positive energy and resourcefulness. I thank Diane Belcher for always encouraging examining issues from multiple perspectives and for steering our applied linguistics ship at GSU. I thank Sarah Carlson for providing inspiration for part of the direction of this dissertation and for letting me know that I could keep moving forward. I would like to thank Xiangying Jiang, Meg Malone, and Alicia Kim for providing me with opportunities to broaden my horizons in research. Each as role models in the field, each in their own way.

I wish to thank everyone in the Applied Linguistics department at GSU who worked with me on various language assessment endeavors. Thanks go to Sally Ren for providing a sounding board for ideas and feedback, moderating the defense, and providing ratings for comprehension

tasks in this study. I would also like to thank Xian Li, Yunjung Nam, Sanghee Kang, Andrew Schneider, and Analynn Bustamante for providing ratings for comprehension tasks.

I also want to extend my thanks to everyone in my Ph.D. cohort who provided feedback on various academic matters and support in times of stress. In particular, Jessica Lian, Katia Vanderbilt, and Selahattin Yilmaz provided moral support and the occasional commiseration. I also want to thank Stephen Skalicky, Cynthia Berger, and Sarah Goodwin for years of advice, and for paving the way before me in the Ph.D. program at GSU.

Critically, this dissertation would not be possible without the financial support from the Adult Literacy Research Center at Georgia State University. Without the Dissertation Support Grant, participant recruitment would not be possible. I appreciate the support and guidance from Iris Feinberg and everyone at ALRC, and I hope that the product of this dissertation does the support grant proud.

I also need to express my deepest gratitude to all the participants who gave their time to participate in data collection for this dissertation.

Finally, I thank my family for always being having my back and providing an unyielding encouragement even when I doubted myself.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AOI	=	Area of Interest
ET	=	Eye-tracking
L1	=	First language
L2	=	Second language
LME	=	Linear mixed effects (model)
MC	=	Multiple-choice
SVT	=	Sentence Verification Task

1 INTRODUCTION

1.1 Background

Second language (L2) students at the post-secondary level face numerous challenges related to language use. Much previous research has focused on L2 writing and speaking while reading development has generally been overlooked as a source of difficulty for learners, compared to writing and oral communication (Andrade, 2009). Yet, reading remains the most critical language skill for academic success at the post-secondary level (Anderson, 1999; Evans et al., 2015; Hartshorn et al., 2017; Jordan, 1997). There are high demands placed on college-level readers regarding how much they must read in a short time, and how they apply the information they read. Importantly, reading is not a language exercise for college readers, but a means to an end; reading is done strategically to learn and engage with the topics they study, and this strategic and purposeful nature is central to academic reading ability (Evans et al., 2015; McNamara & Magliano, 2009; Moss et al., 2011). The ability to read for these purposes in a second language involves not only decoding skill, but also involves higher order comprehension of the meaning of texts. When a reader's purpose is reading to learn, the focus is "to construct an organized representation of the text that includes major points and supporting details" (Enright et al., 2000, 4), and any measurement of advanced L2 reading comprehension needs to activate higher-order processing, global text reading, and reliance on more than surface-level linguistic features (J. C. Alderson, 2000; Enright et al., 2000; W. Grabe, 2009).

Yet L2 reading assessment methods purported to measure comprehension may not target critical higher order processing skills reflective of academic reading-to-learn demands placed on L2 readers. The lack of investigation into higher-order reading processes during L2 reading assessment may stem from the notion that L2 reading ability is often considered to be primarily

comprised of L1 reading skill and L2 oral proficiency (Koda, 1988, 1990), with little attention to how reading subskills develop or redevelop in advanced L2 readers. L2 reading comprehension assessments have thusly relied primarily on examinees' responses to practically-scored, discrete, closed-ended items, such as multiple-choice questions, regarding information from a text (Daza & Suzuki, 2004; Enright et al., 2000). However, it is well-established that during text comprehension, the types of texts and tasks (i.e. an activity in which information from a text is put to use) activate various comprehension processes and strategies (Kaakinen & Hyona, 2005; Kamhi & Catts, 2017; Miller, McCardle, Cutting, & Dyslexia Foundation, 2013; Moss, Schunn, Schneider, McNamara, & VanLehn, 2011; Ozuru, Briner, Kurby, & McNamara, 2013). Reliance on a discrete, closed-ended task to assess L2 reading, like answering multiple-choice items, warps the reading construct by providing assumed choices, leading readers to look for linguistic cues in answers and use test-wiseness strategies over reading strategies (Rupp et al., 2006).

Alternative measurements of comprehension which elicit an individual examinee's representation of a text have been employed previously, such as cloze tests (Carrell, Carson, & Zhe, 1993; Williams, Ari, & Santamaria, 2011) and summary tasks (Enright et al., 2000; Seidlhofer, 1990), but the required production component for these reading comprehension assessments creates construct-irrelevant variance as well. Further, although cloze and summary tasks arguably tap more directly into readers' mental models, little is known about the nature of these open-ended reading tasks differing from closed-ended tasks in terms of higher-order processing and text-reading behavior.

From the perspective of higher-order text processing, comprehension of a text involves the construction of mental representations integrating content from the text with the reader's own interpretations and background knowledge (i.e., a situation model; Broek, Bohn-Gettler,

Kendeou, Carlson, & White, 2011; Dijk & Kintsch, 1983; Horiba, Broek, & Fletcher, 1993; C. A. Perfetti, 1997; Zwaan & Radvansky, 1998). The construction of such a representation includes the ability to make *inferences*, or to fill in ideas between and beyond the lines of text, necessary to create a situation model (Cain, Oakhill, Barnes, & Bryant, 2001; Carlson, Seipel, & McMaster, 2014; Irmer, 2011). In L1 reading research, there is an increasingly clear picture of the contribution of inferencing ability to reading comprehension (Cain et al., 2001; Carlson et al., 2014; L. Taylor, 2013; Zwaan, 2016; Zwaan & Radvansky, 1998). However, less is known about the contribution of inferencing skill in reading comprehension for L2 readers and whether commonly used second language reading assessment tasks, such as multiple-choice question tasks or summarizing tasks, tap into inferencing skill.

How a reader reads a text in terms of reading rate and attention across a text has also been researched in L1 and L2 reading comprehension (Berzak et al., 2018; Carver, 1997; Enright et al., 2000). These variables in online reading behaviors can be investigated using eye movement data (Conklin et al., 2018; Rayner, 1978; Rayner et al., 1980). Such data gathered from eye-tracking methods presume that eye movements and fixations relate to attention (Marcel A. Just et al., 1982). The insight gained from eye-tracking regarding lexical and syntactic processing during reading comprehension is well attested (Clifton et al., 2016), but less research has investigated eye movement behavior using larger text components as units of eye-movement measurement (Conklin et al., 2018; Kaakinen & Hyona, 2005).

Few studies have examined L2 reading assessment and eye-movement. Bax (2013) compared cognitive effort between areas within texts, and McCray and Brunfaut (2018) compared reading behavior across different levels of texts. Studies thus far have not compared online eye movement behavior between different L2 reading tasks used to assess text

comprehension, let alone in realistic tasks where reading and task completion can be performed synchronously. Thus, there is a need to better understand how types of L2 reading comprehension assessment tasks differ in examinee reading behavior during reading and task response.

1.2 Purpose of the study

Reading in an additional language (henceforth L2 reading) involves a mix of low-level linguistic processes and higher-order comprehension processes placing demands on multilingual readers to read strategically and purposefully. However, understanding of the involvement of higher-order processes in language learners' successful comprehension during assessment remains unclear. In addition, it is unclear to what degree different assessment tasks (e.g. MC questions, cloze tasks, and summary writing) elicit higher-order processing.

This dissertation presents an effort to further understand reading processes activated during a realistic L2 reading comprehension assessment situation. It is not possible to gather direct information about the internal mechanisms of higher-order processing during a realistic assessment scenario. However, components of higher-order processing such as inferencing or selective attention can be operationalized using various online measures, such as eye-tracking, and offline measures, such as post-hoc tasks which tap into activation of processes primed by stimuli. Although higher-order mental modeling strategies, such as inferencing, rereading and integration of information across pieces of texts, are known to influence comprehension in monolingual readers who already have developed language proficiency, less research has focused on how these abilities contribute to comprehension for adult multilingual readers and language learners. Because academic literacy skills may continue to develop alongside general proficiency skills in their L2, the focus on reading in multilingual research has been

predominantly on proficiency. It is unclear whether or not higher-order reading processes, which play a role in monolingual readers' comprehension, contribute to reading comprehension for multilingual readers. This lack of understanding poses a threat to L2 reading assessment validity. Bachman and Palmer (1996), in their test-authenticity argument, state that use of a language test is justified when we can "demonstrate that performance on language tests corresponds to language us in specific domains other than the language test itself" (p. 23). In this way, if it cannot be established that a reading test does not activate the processes required by the target reading domain, it cannot be considered valid. Thus, several assessment tools were examined in this study to determine reading task validity in measuring higher-order comprehension skills for multilingual readers.

In this dissertation, three tasks were used as measures of reading comprehension: multiple-choice questions, cloze tasks, and summary tasks. These tasks were chosen because they represent the different levels of constraint and construction which differentiate realistic L2 reading assessment tasks. Completion of these tasks was analyzed under the lenses of higher-order processing and text-reading behavior in terms of representative constructs. To understand L2 reading comprehension in terms of higher-order processing, inference activation was compared between the above tasks, and the relationship between inference activation and reading comprehension scores were analyzed. To understand reading task performance in terms of reading behavior, task differences were also examined using eye-movement behavior variables, as well as comparison with score (described with more specificity in the methods section). To predict scores, statistical modeling of scores was carried out using inferencing and eye-movement metrics, including predictor individual difference variables: L2 English proficiency, reading speed, working memory, reasoning, and motivation). This research will help the field of

reading comprehension assessment further understand the cognitive and construct validity of these assessment tasks.

1.3 Research Questions

The purpose of this dissertation is a novel investigation of two cognitive process domains important to reading comprehension: inference making and eye-movement behavior. Data from a reaction time paradigm task and data from eye-tracking methods were analyzed and compared to performance on reading comprehension outcomes on three different reading tasks. Additional comparisons are made to baseline individual differences which influence L2 English reading: Morpho-syntactic proficiency, logical reasoning, working memory, reading speed, and motivation. The goal is to better understand the L2 reading process from a cognitive perspective and understand how these processes can be measured during L2 reading comprehension assessment. This involved two major lines of inquiry.

The first is understanding inference generation across reading tasks (i.e., responding to MC questions, cloze tasks, and summary tasks). To address this avenue of study, L2 English readers were asked to complete the three mentioned comprehension tasks for three different reading passages (one each). Activation of inferences made during reading was measured using a post-hoc, sentence verification task after each of the three comprehension tasks. In this task, participants responded as quickly as possible to a series of sentences with a true or false response. Inference activation was operationalized as reaction times to sentences which contain information inferable from, but not occurring in, the text from the comprehension task. This is discussed further in the methods (chapter 3). This line of inquiry involves the following sub-questions:

1a. Do examinees respond significantly faster to sentences inferable from a text than to unrelated sentences after reading the text and is this affected by the type of reading comprehension task?

1b. To what extent does inference generation predict variance in comprehension task outcomes (scores) beyond variance predicted by individual differences in proficiency, reasoning, memory, reading speed, and motivation?

The second line of inquiry is to understand if comprehension scores are related to online reading behaviors (i.e. eye-tracking). The goal of this line of inquiry was to understand how online reading behavior differs between multiple-choice, cloze, and summary test items, and whether differences in reading behavior contribute to an examinee's reading comprehension performance on these tasks in a meaningful way. A variety of eye-tracking metrics were gathered, and they are discussed more thoroughly in the following two chapters. This inquiry includes the following sub-questions:

2a. To what extent does online reading behavior, as measured by eye-tracking, differ between reading tasks?

2b. To what extent do online reading behaviors predict variance in reading comprehension scores beyond that predicted by individual differences?

2 LITERATURE REVIEW

This section contains a review of previous literature on the subject of L2 academic reading comprehension. This includes a) a review of models of text comprehension and higher-order reading processes, b) a review of research on the use of the three L2 reading comprehension tasks examined in this dissertation, and c) a review of literature examining text-level reading vis-à-vis inferencing and eye-movement behavior. This chapter concludes with a return to the purpose of the current study, to investigate cognitive processes occurring during second language reading assessment, presenting hypotheses for expected findings based on previous literature. Although many of the theories and practices regarding assessment of reading in an additional language may apply to multiple language situations, research in this field is primarily focused on English as a Second Language, so findings from previous studies and the framing of the current study are somewhat shaped by the prevalence of English language testing.

2.1 Components of reading comprehension

The validity of a reading comprehension assessment is dependent on the underlying model of the reading process upon which an assessment is designed. The interpretation of reading test scores, the types of tasks utilized, the content of reading passages included, and the target of individual test items all depend on test creators understanding of the component skills and processes which constitute reading comprehension. This section reviews theories of the components of reading comprehension, beginning with a review of monolingual reading comprehension and ending with the additional complexity in reading in an additional language.

2.1.1 Monolingual (L1) modeling of reading comprehension

Reading comprehension is considered not a single uniform construct, but rather a conglomeration of psychological and linguistic processes which contribute to understanding the

language and ideas found in text (Kintsch & Yarbrough, 1982). Despite impressions of reading comprehension as a receptive process, comprehension is considered to be a constructive process, as information in a text does not give itself to a reading, but the reader must actively extract information and build a model of the information (Snow, 2002). This successful construction is built upon the activation of many interworking processes, including lower-order skills which are used to construct meaning from the bottom up (decoding, activating vocabulary, identifying local syntactic/cohesive cues) and higher-order skills to construct meaning from the top down (activating schemata, inferencing, strategy use) (Afflerbach, 2016; Van Dijk & Kintsch, 1983). The comprehension processes are further influenced by who the reader is, for what purpose they are reading the text, and how they distribute their attention throughout a text and strategically activate various processes (Afflerbach, 2016; Grabe, 2009; Khalifa & Weir, 2009; Urquhart & Weir, 2014).

2.1.1.1 Lower-order reading skills and bottom-up processing

In general, reading is considered to consist of both more basic, *lower-order* processes and more cognitively complex, *higher-order* processes. Lower-order processes include grapho-phonemic processing (i.e. making sound-symbol correspondences), morphological awareness, word recognition, syntactic parsing, and local activation of semantic knowledge. From a receptive skills perspective, the key aspect of lower-level reading processes is word recognition, with each lower-level process facilitating the goal of recognizing the words on the page (or screen) (Perfetti, 2007). Bottom-up reading processes are relatively linear, and the content extracted during lower-order stages of reading are considered relatively stable across individuals with similar skills (Bernhardt, 2011).

2.1.1.2 Higher-order reading skills and top-down processing

Higher-order processes, on the other hand, include making inferences between referents in a text, activating background knowledge, and evaluating the purpose of texts and the usefulness of information. Reading from the top down allows entails more flexibility across individuals in the ultimate interpretation of a text based on readers' backgrounds and purposes. In a sense, this is where meaning from the text is constructed by the reader. Readers use the literal text as cues to activate connections between propositions using their logic and inferencing skills, background information and experience with previous texts. Higher-level processing is seen as having two levels (Kintsch, 1998; Grabe, 2009): a text base comprehension level, where a reader creates a model of ideas and propositional content found in a text, and a situation model level, where the overall meaning of a text is constructed by the reader through connecting propositions and relating content to background knowledge and reading context.

Models of reading often emphasize the integration of higher-order and lower-order skills when constructing comprehension. The Construction-Integration Model (Kintsch, 1998; Van Dijk & Kintsch, 1983) suggests successful comprehension ultimately rests on both successful decoding of a text base and successful construction of a mental model. Alternatively, interactive approaches to reading the assert that reading deficiency in one aspect of comprehension can be compensated by strengths in another aspect (Stanovich, 1980). For instance, lack of knowledge of a particular lexical item can be compensated by stronger inferencing skill so understanding can be maintained.

2.1.2 Factors which influence reading comprehension

Despite the frameworks of reading above, reading comprehension is not an isolated skill which is simply the sum of its parts, e.g. decoding, text modeling, and mental modeling.

Successful reading ability is also influenced by numerous other cognitive and non-cognitive factors. This include factors from print exposure to general knowledge to metacognitive awareness. An exhaustive list of factors is outside the scope of this dissertation, but a few important factors are mentioned below.

Comprehension is impacted by processing efficiency; i.e. how fast someone can take in visual information from a text. The speed of decoding words and the efficiency of processing of visual information has been found to correlate strongly with comprehension of texts (Artelt et al., 2001). Faster processing allows for more information to be accessible in short-term memory and frees up cognitive capacity for higher-order skills.

Along the same lines working memory capacity is itself also a factor of successful comprehension (Daneman & Merikle, 1996). Working memory capacity allows for better temporary retention of text information which can be updated with new information. This can contribute to stronger mental modeling. Another cognitive factor which impacts reading comprehension is logical reasoning. This has been shown to relate to comprehension, specifically to the way readers connect pieces of information across a text (Segers & Verhoeven, 2016). Specifically, reasoning ability is critical for making inferences.

Finally, non-cognitive factors may also impact reading comprehension. Motivation to read has a strong impact of literacy outcomes. Motivation is typically divided into extrinsic motivation, which comes from external material and social influences, and intrinsic motivation, which is more related to genuine interest in an activity. In L1 reading contexts, higher motivation has been found to predict positive reading development (Guthrie et al., 2007). It is posited that intrinsic motivation is especially critical for reading development (Csikszentmihalyi, 1990).

2.1.3 Influences on multilingual reading

Up to this point, this discussion of reading comprehension has been rather neutral regarding whether reading is done in a first or second language. Unlike with L1 reading comprehension development, most L2 readers become L2 comprehenders well after L1 reading comprehension skills are developed (Jiang, 2011; Koda, 1988). Further, reading in a second language involves higher cognitive load than reading in an L1, as language processes which are assumed to be fully automatized in L1 reading may still be developing in L2 reading (Yoshida, 2012). Alderson and Urquhart (1984) summarize the conflicting hypotheses about what influences reading in an additional language as follows:

1. Readers who are competent readers in a first language will be competent readers in an additional language.
2. Successful reading in an additional language is a product of knowledge or proficiency of the additional language.
3. Poor reading in an additional language is due to lack of application of relevant L1 literacy skills. This supposes that there is a threshold of language ability before literacy skills can be applied to reading. Below the threshold, the cognitive demand of using a second language is too high for L1 literacy skills to be utilized.
4. Poor reading in an additional language is due to a mismatch of literacy skills in the first and additional language, i.e. multilingual readers *do* apply known literacy skills, but they may not aid reading an additional language.

Field (2018) generalizes this further, identifying the two modern lines of argument being a universalist argument which posits all readers at some point achieve a set of literacy skills that contribute to comprehension on the one hand, and an expertise argument which posits that there

is a language proficiency threshold which must be reached before literacy skills can be engaged on the other.

The earliest position on multilingual reading was that deficiency in reading in an additional language was a result of poor literacy in the L1. This notion rose from the idea that reading in any language involves the same set of strategies (Goodman, 1973, in Alderson & Urquhart, 1984) and argued that multilingual reading instruction involved rectification of poor L1 reading habits (Coady, 1978, in Alderson & Urquhart, 1984). This view was supported by correlations found between success on L1 and L2 basic aptitude measures and cloze tests. However, little evidence has been produced beyond bidirectional relationships, and there has been little empirical support for the hypothesis that reading ability in an additional language *is* reading ability in the first language.

Evidence seemed to be found more readily for the second hypothesis, that reading in a second language was dependent on second language proficiency. The aspect of proficiency could be related to vocabulary, i.e. knowing the words needed to represent concepts in a text (Ulijn & Kempen, 1976) , or be related to more general L2 proficiency (Cziko, 1978). These studies showed that the correlation between L2 proficiency and L2 reading was higher than that between L1 and L2 reading, yet these studies also often found moderate correlations between literacy in both languages (Alderson & Urquhart, 1984).

More likely, there is a complex interaction of language proficiency and L1 literacy skills in reading in an additional language. This is the stance put forth by the threshold hypothesis, which implies that once readers reach a certain threshold of L2 proficiency, L1 reading skills can be applied, and that both are necessary for reading comprehension in an additional language (Cummins, 1979). In recent years, researchers have agreed that there is likely a mix of influences

on L2 reading from the L1 and L2, investigating how much positive transfer of literacy skills exist in developing L2 reading comprehension, and what the unique contribution of L1 reading skills and L2 language proficiency are for L2 reading. Mokhtari and Reichard (2004) found that the consciously activated reading strategies of multilingual English readers did not differ from monolingual English readers in comprehension of texts, although the contribution of strategies to success may differ, supporting the view that literacy skills are shared between good comprehenders in an L1 or an L2, and L2 readers can transfer their literacy skills from their L1. At the same time, it is well established that there are measurable linguistic thresholds to comprehension, such as the need to comprehend 95% of the vocabulary of a text to achieve minimum comprehension (Laufer & Nation, 1999; Verhoeven et al., 2011) and the role morphological awareness plays in L2 text comprehension (Nagy et al., 2006).

Indeed, most studies examining this issue have found that each domain contributes meaningfully, but not overwhelmingly, to L2 reading ability (Carrell, 1991; Carson, Carrell, Silberstein, Kroll, & Kuehn, 1990; Jiang, 2011; Pae, 2017). In more cognitively challenging reading tasks, which may be less familiar to readers in L1 or L2, the difference between contributions of L2 proficiency and L1 reading skills widened, and L2 proficiency takes the lion's share of predictive power for L2 reading. L2 proficiency also influences the way in which readers arrive at comprehension, as text coherence is based on different cues for speakers across proficiency levels, with lower proficiency readers attributing coherence to semantic similarity throughout at text and higher proficiency readers attributing coherence to causal linkage throughout at text (Nahatame, 2014). Pae (2017), in modeling the componentiality of L2 reading as a combination of L2 proficiency and L1 reading skills, found that both aspects contributed to L2 reading, with L2 proficiency being the stronger predictor of L2 reading ability, but the

strength of contribution differed depending on the cognitive demands of the task. In more cognitively challenging reading tasks, the difference between contributions of L2 proficiency and L1 reading skills widened. Pae (2017) offers no explanation for why the gap in contribution should widen, but it may be that more cognitively complex tasks begin to involve more register specific language that has no analog in the L1, and thus L2 proficiency takes the lion's share of predictive power for L2 reading.

This calls back to the fourth hypothesis mentioned by Alderson and Urquhart (1984), that successful reading in a second language depends on learning skills and strategies specific to the language. This hypothesis rests on the idea that every language's text conventions require certain literacy skills that may not be present in all languages and reading instruction and assessment should focus on second language literacy skills as distinct from either proficiency or monolingual literacy. This hypothesis has its roots in outdated contrastive analysis (Cowan, 1976), focusing on the misapplication of L1-specific reading strategies.

A more modern synthesis of this hypothesis highlights the importance of literacy strategies but diminishes the labeling of them as L1 strategies. This can account for the fact that in a globalized world, academic systems often encourage use of academic literacy skills in an additional language beyond that acquired in first languages. For at least English, large populations of learners come from language backgrounds lacking in strong emphasis on print literacy (Bigelow & Tarone, 2004), and many learners are early bilinguals which only develop literacy in one language or another (Ramírez, 2000). The skills needed for academic reading with these reader populations may be distinct from both oral L2 proficiency and presumed L1 literacy skills.

Reading comprehension assessments for multilingual readers are designed as proof that a reader holds the necessary skills for comprehension, be they related to language proficiency or literacy skills (Green, 2013). An assessment use argument for a test of advanced level reading for academic purposes must be able to attribute real-world ability to reading comprehension scores. as seen in standardized proficiency tests used for university entrance, scores on reading comprehension tests or subtests can be considered valid only if they reflect test takers' various capabilities to comprehend texts in realistic situations reflective of college-level academic reading in an additional language. This implies reading test design for multilingual readers cannot be identical to L1 reading tests, nor can it focus overly on language features simply presented as reading exercises. The assessments must utilize texts which are general enough so as to tap into various knowledge domains without over-emphasizing the role of any specific content. Reading assessments must evaluate skills from the bottom up and the top down to make the claim that a reader is ready for demands of academic reading, which is dynamic and involves multiple reading purposes.

2.1.4 Highlighting inferencing

The ability to make inferences is a critical higher-order comprehension skill. Inferences are the implicit pieces of information a reader creates to go beyond the explicit propositions of a text and link ideas from a text to each other, to background knowledge, and to predictions about text (Cain et al., 2001). Inferences take place at the word level, when meaning of an unknown word is inferred. More critically, readers make inferences to connect new ideas to prior knowledge to support text understanding, and resolve connections not explicitly made between textual propositions (Khalifa & Weir, 2009). Readers must make some number of inferences while reading, but it has been posited that only as few inferences generated are needed (McKoon

& Ratcliff, 1992; Ridgway, 1994). Just & Carpenter (1987) made the distinction between backwards and forwards inferences. The following example illustrates backwards inferencing:

When she finished reading the letter, she watched its pieces burn away in the fireplace.

Beyond understanding individual words and phrases, you must understand the two propositions depicted in the scene and understand the chronological link between the two propositions. You may also make the inferences that, "She tore the letter into pieces," and, "She threw the letter into the flame," which are not overtly expressed in the text base. Additionally, the necessity to generate each of those inferences varies, with, "She tore the letter," less required for adequate comprehension of the sentence and "She threw the letter into the flame" more required.

Forward inferences are similar, but rest on the reader making a connection between a proposition in the text to a reader's prediction of a future text model. Reading assessment researchers argue that reading comprehension assessments should emulate the real-life aspects of making inferences regarding vocabulary and inferences to general knowledge to the extent that it is fair for the various test-takers' backgrounds but must especially focus on inferences which make connections forwards and backwards within a text.

2.1.5 Real-time reading behavior

Reading comprehension is also moderated by the real-time behavior a reader engages in. The way in which a reader engages with a text and where they spread their attention is based on reading purpose, strategic decisions made by the reader to facilitate comprehension, and the reader's understanding of schemata which inform them about where to look to extract important information.

Urquhart & Weir (2014) emphasize the importance of goal setting in determining reading behavior. Each reader sets specific goals for comprehension based on the purpose for which a text is to be comprehended. This is especially critical in reading assessment, where the assessment task sets a purpose for the reader, and they must tailor their goals to the purpose. Setting goals additionally influences the relative importance of lower-order and higher-order skills during reading.

Real time reading behavior can be either *local* or *global* depending on reading goals. *Local* reading entails focus of attention on specific local text regions, ad implies an emphasis on lower-order skills. Reading to understand specific words and sentences or reading to find an explicitly stated fact occur at the local level. Many reading questions on tests of overall language proficiency involve items which activate local reading (Enright et al., 2000), with questions that can be answered by finding linguistic connections between the question and discrete pieces of text (J. C. Alderson, 2000).

Global reading involves more higher-order processes. This entails comprehension of the macro-structure of the text (Kintsch, 1998). Understanding the macro-structure allows readers to comprehend the main ideas, or gist, of a text, and allows for an index of locations of propositions within a text for quicker access. Global reading is activated when readers build a model of a text's main idea, skim for gist, or search for specific ideas during rereading.

Reading can also be *careful* or *expeditious*. These reading behaviors relate closely to the rate of reading and relative level of attention paid to the language and propositions in a given region of text. Careful reading is enacted to fully comprehend a text. This could happen at the local level when lower-order decoding is called for, and at the global level when a reader is reading to learn the content of an entire passage. This type of reading is typically slower and

incremental (Rayner et al., 2006). Careful global reading is especially relevant in academic reading, where thorough understanding of academic expository texts is demanded. It entails the complete building of a mental model. Tests such as the Cambridge English Proficiency Exam specifically seek to evaluate reading at this level (Khalifa & Weir, 2009).

Expeditious reading occurs at a faster reading rate. This could happen when a reader has a high enough level of lower-order skill proficiency to decode rapidly and enough schematic knowledge of a text to rapidly construct a mental model (Carver, 1997). This efficient reading is seen as the goal for reading development (Khalifa & Weir, 2009). Expeditious reading also includes compensatory reading strategies commonly taught to learners who have not reached the level of efficiency to carefully understand text quickly, such as skimming, searching and scanning. Skimming is the most global expeditious reading strategy, where reading is done rapidly, with decoding done at sampled sections of text so that the reader can form the gist of a text while ignoring minor details. Scanning is more local, with quick linear eye-movements made across a text until specific relevant lexical items are decoded, upon which local careful reading occurs to comprehend details. Search reading occurs somewhere in between, where a reader makes an attempt to quickly identify locally readable information but activates global knowledge of the text structure to quickly identify the location of the information. Khalifa and Weir (2009) point out that reading assessment often focuses on comprehension at the global, careful level, but due to test constraints (such as time limits), encourage expeditious reading.

2.2 Reading comprehension tasks in assessment of academic L2 English proficiency

2.2.1 Reading purpose and task design

Successful academic reading involves efficiency of processing, language knowledge, strategy use, print exposure and background knowledge, working memory, and metacognitive awareness of reading goals and purposes (Grabe, 2009). These skills are activated in accordance with a reader's purpose. Academic reading involves engaging in different types of reading to fulfil certain academic purposes. These include, roughly in order of fastest and least cognitively demanding to slowest and most cognitively demanding, reading to search out specific information (*scanning*), reading for quick understanding (*skimming*), reading for general comprehension, reading to learn, reading to integrate information, reading to evaluate or critique, and reading to memorize (Carver, 1997; W. Grabe, 2009). Reading assessment can focus on any one of these purposes, but must necessarily prioritize activation of some skills and processes over others (W. Grabe, 2009; Urquhart & Weir, 2014).

Based on the constructs and purposes underlying reading in an additional language, multilingual language reading assessment has been aimed at measuring different aspects of the reading process. Reading assessments can include questions which target vocabulary knowledge and propositional knowledge on the lower-order side. This is used in both lower-level L2 achievement test and general L2 proficiency tests (J. C. Alderson, 2000; Enright et al., 2000; Genesee & Upshur, 1996a). However, advanced academic reading tests have focused particularly on assessing skills related to understanding of a wide range of text types, comprehending main ideas and details with texts, identifying important ideas, and differentiating fact from opinion (Khalifa & Weir, 2009). At the level of text comprehension, reading comprehension assessment mostly focuses on global and higher-order comprehension. Academic reading tests typically ask

readers to identify pieces of information which may be implicit, such as an inferred connection between propositions or the intent of an author in including certain information (L. F. Bachman, 2000; W. Grabe, 2009; Khalifa & Weir, 2009).

2.2.2 Reading comprehension task types

The type of task used in an assessment can drastically affect the way reading comprehension skills are activated. Reading assessors must be aware in how task types draw on these different interlingual competencies. Since there is no one-size-fits-all task for assessing general reading comprehension, a plurality of answer formats is the best way to build a picture of reader comprehension in reading assessment situations (Alderson, 2000; Grabe, 2009).

Assessment research has consistently acknowledged the effect of task type and response format on what aspects of comprehension is measured through reading assessment (J. C. Alderson, 2000; Brantmeier, 2005; Grabe, 2009; In'nami & Koizumi, 2009; Khalifa & Weir, 2009). The more a task relies on skills not related to comprehension (i.e., noticing verbatim overlap in a multiple-choice question, possessing strong writing skills), the less a task can be used as a valid measurement of reading comprehension (J. C. Alderson, 2000; Lee, 2011). This can be the case when a measure of reading relies too much on the surface structure of a text, allowing for grammatical cues (structural overlap between a test item and text, answer choices which can be ruled out due to language errors, etc.). Researchers are interested in the differences between selected response formats (e.g., multiple choice, or MC, questions), open-ended discrete response formats (e.g. cloze items), and constructed response formats (e.g. short answer questions, summary-writing) to measure comprehension of a text. Assessment users must be more aware of the trade-off between using narrow and practicable assessment items and having a

holistic picture of reading comprehension ability; the fewer types of items used, the narrower the validity argument for comprehension assessment must be (Lyle F. Bachman, 2002).

Discrete multiple-choice items, for instance, give assessors a chance to target specific text base and situation model features, eliciting test-taker knowledge about literal, surface-apparent facts as well as inferable propositions. However, they also provide readers an unintended crutch with subtle extraneous information which inhibits activation of situation model building. The mere construction of a discrete, closed-ended question shows that most of the situation-model-building leg-work has already been done by the test designer, and test-takers primarily need to activate problem-solving strategies to match item-writer's comprehension of the text; an inferential process, to be sure, but not tied enough to textual inferencing. Open-ended tasks allow for the activation of the important higher-level reading comprehension processes of inferencing, accessing pragmatic competence, and activating background knowledge. Test users worry, however, that open-ended formats a) allow the test-taker to leave out aspects of comprehension because they were not explicitly elicited, or b) bog down the test-taker with use of extraneous, construct-irrelevant parsing, writing, and editing skills.

O'Reilly and colleagues (2018) found it was not specific tasks that influenced reading behaviors, but the way tasks encouraged readers to set goals before reading. Utilizing visual inspection of eye-tracking data, they found that when readers were given an explicit goal to achieve from reading, higher-order reading processes were elicited. Without the goal setting, readers were more likely to perform quicker text reading, opting to compare MC questions, the only available motivator for reading, to segments of text. Providing readers with an overarching goal induced more careful first reads. From this, it may be the case that one specific reading task

does not motivate specific fundamental reading behaviors, but rather how the task is presented to readers as a goal that determines the readers approach to comprehension.

The discussion so far has used the notion of task very broadly to refer to the method test designers require test takers to demonstrate understanding of a target text. There are numerous types of tasks for assessing reading comprehension in a second language, and each one entails multiple dimensions of variation. An exhaustive description of the different types of tasks used in reading comprehension assessment is beyond the scope of this dissertation, but a few common task types will be detailed. This section reviews three common tasks utilized in language tests to assess L2 reading comprehension. Few formats of L2 reading comprehension assessment, in ESL or otherwise, are unique to the second language testing sphere, and these tasks are often utilized in general literacy research and assessment. The format of comprehension assessment tasks can be either *selected-response,* where the reader choses from a discrete number of answer choices written by the test designer, or *constructed-response,* where the reader must produce an answer choice. Construct-response tasks further differ in whether the answer is *closed-ended,* with a specific object correct answer expected by scorers, or *open-ended*, with more production expected from the reader which is graded more subjectively. Three tasks have been chosen to represent selected-response tasks, closed-ended constructed-response tasks, and open-ended constructed-response tasks. They are multiple-choice question answering tasks, the cloze task, and the summary task. Each of these is described below.

2.2.2.1 Multiple-choice and other selected-response tasks.

The multiple-choice question format and selected-response formats in general are very versatile, as questions can be formulated to target vocabulary knowledge, understanding of main ideas, understanding of subordinate details, comprehension of implied information, predictions

about text purpose or subsequent readings, and supposition about author intentions and opinions. The questions put no language production demands on test takers, and scoring decisions are practical and objective. The responsibility is on the test designer to ensure that questions and correct options truly tap into the intended construct and do so in a way that is fair to test takers from various cultural backgrounds. Discrete item formats include multiple-choice questions, true/false questions related to a passage, fill-in-the-blank items with a word bank, or even more complex tasks, such as selecting a sentence to complete a paragraph and text reordering tasks. This type of task is very prevalent in assessing L2 reading comprehension, and is a major tool for measuring L2 reading comprehension on tests such as the Test of English as a Foreign Language (TOEFL; Enright et al., 2000), the Main Suite Cambridge ESOL examinations (Khalifa & Weir, 2009), or the Test of English for International Communication, or TOEIC (where it is the only type of item; Daza & Suzuki, 2004).

When used to measure reading comprehension, multiple-choice (MC) items typically involve a question stem which is answered by selecting from three or more possible pre-written options, of which a subset are correct options, or *keys*. The preference for MC items typically stems from MC items' requiring no production from the examinee, seemingly reflective of the receptive nature of reading (Genesee & Upshur, 1996a), and being practical to administer and rate (Khalifa & Weir, 2009). However, the separation of text information into discrete units identified by the test designer, each with objectively correct or incorrect options, implies that a singular correct reading and modeling of a text exists, which may not be the case for all texts. The foreknowledge that a keyed answer exists allows examinees to view MC items as problem-solving tasks, requiring discrete use of surface strategies rather than global comprehension processes, even when the questions may attempt to target implicit information or general gist

(Alderson, 2000; Daza & Suzuki, 2004). Rupp, Ferne, and Choi (2006) used think-aloud protocols to classify examinee strategies when addressing multiple-choice reading comprehension items and found that the strategies utilized were closer to lower-level processing and problem-solving strategies than higher-level meaning construction strategies. Thus, additional care must be taken in designing MC items which target a variety of reading processes and do not relate to closely to isolatable lexical items in the text or superficial details to avoid having the MC task rely off-construct reader abilities. At their best, MC questions can target a variety of abilities and be used for multiple performance and diagnostic purposes, but require careful construction on the part of the test designer (see for example, Carlson et al., 2014).

2.2.2.2 Cloze and other closed-ended tasks

The first major alternative to the discrete selected-response item format is a format which requires some constructed response in the way understanding of text is demonstrated by the reader. This includes fill-in-the blank statements related to a passage (with no word bank), diagram labelling tasks, or test re-construction tasks like the cloze or c-test. This format is useful for providing response flexibility without imposing too many linguistic demands on the test taker and removes some of the threats to the validity of selected-response items by removing some of the superficial cues to the correct answer. However, these are often disfavored because the minimal flexibility provided comes with a severe drop in practicality. However, the cloze task in particular is purported to cover complete text understanding and is simple to construct, even if scoring is less practical than in selected-response formats.

The cloze test task is a specific type of fill-in-the-blank item which allows for examinees' individual input, making them more open-ended, while still having narrow expectations on what responses are allowed. Cloze test design involves deleting words in an otherwise coherent text

and replacing the words with blanks that examinees must fill in with an appropriate word (or sometimes phrase). In some versions, choices are provided, and in others, test takers must provide their own words. The C-test provides an area in between providing and hiding answer choices by having the first half of the target word already present in a blank. Cloze tests and C-tests have variably been used as reading comprehension tests and as a general proficiency test, which raises concern about the validity of such a test to measure any second language skill domain, such as reading, in isolation (Alderson, 2000). It additionally presents reading material in an artificial manner, which may not reflect realistic academic reading purposes.

Variations on cloze testing can be used to home in on semantic content of a text to elicit reading skills more specifically (Carrell, 1993). Random or systematic-deletion cloze tests with interval deletion of words may target general second language proficiency or syntactic knowledge, as they require test-takers to activate a broad base of language proficiency dimensions depending on what is deleted in terms of part of speech and function words. However, rational-deletion cloze tests can better target semantic content of texts (Kleijn, 2018) and the logical connection of information in a text (Greene, 2001). There is also the issue of objective scoring. Although cloze tests are meant to be objective test tasks, deleting on a regular interval can lead to blanks where multiple possible right answers exist. This opens the test up to invalid interpretations in the case where only the expected response is accepted, or makes for less practical rating, especially for an objective test.

2.2.2.3 Summary and constructed-response tasks.

Various item formats for assessing text understanding rely more on reader production. These items seek to further extend the flexibility of response, giving the reader more freedom to present their own understanding of a text, at the further expense of practicality. The hope is that

by allowing freer reader production, a deeper sense of the reader's understanding can be elicited. The simplest constructed-response format is the short-answer questions format, which is similar to MC questions, but requires a free response from the reader instead of selection from options. Construct-response items could also be as complex as composing a position paper based on the reading of a text. Despite the additional level of impracticality, constructed response tasks like short answer questions about a passage encourage the construction of text models and higher-order text integration in a way a constricted response format cannot.

Perhaps the most direct subset of reading comprehension assessments are those which have readers report what they learned from or about a text. In its rawest form, this type of assessment appears as a recall task, with readers explicating the propositions they remember from a text. A more nuanced task of this type is a summary task, which demands more directed, purposeful text modeling than direct recall. Summary tasks, as reading comprehension assessments, are productive tasks where the examinee is asked to produce a condensed report of the content in a reading passage which is evaluated for accuracy and detail. The summary task relies less on writing ability than the more conceptually demanding task of integrated reading-writing as found in source-based essays. However, summary tasks still cede more control to the reader in modeling the text and preparing a response than short answer questions which rely on item designers' mental models similarly to MC items. The need for production by the reader adds a layer of, sometimes unwanted, difficulty to the response process, but can also be seen as more solid evidence of understanding discourse structure (Spivey, 1990). Ji (2011) confirmed that written summaries rely too heavily on writing, and are not suitable tasks for lower-level examinees. It is thus important to understand what aspect of a summary needs to be evaluated to assess reading comprehension. Benzer et al. (2016) found that summary writers with better

comprehension of a text write shorter, quicker summaries with less direct quotation. These results highlight the fact that quality is more important than quantity in using productive tasks to assess reading comprehension skills; text length and direct, keyable items, like borrowed text, may not be useful for rating summaries. Wang and colleagues (2017) looked at the influence summary writing had on reading behavior in reading comprehension testing, finding that the summary task elicited longer reading times from readers than a MC question-only task, and that less efficient readers benefited from longer reading times. This highlights the fact that task types may induce different reading behaviors in test takers.

There are many gaps in the research on summary as L2 reading comprehension assessment. Few studies have examined the summary writing of advanced academic L2 readers reading in the academic target language-use (TLU) domain. There is also no research which used a summary rating method that controlled for rater judgments of writing quality in assessing summary accuracy and text modeling. Considering that Moss and colleagues (2011) found self-explanation (McNamara, 2004) to be useful aid in comprehension, and that authentic academic reading relies on the reader's autonomy in constructing a mental model, without the crutch of another's (e.g. a test item writer's) cues or assumptions to guide them, it is worth exploring summary assessment as a reading task reflective of real-world reading to be used in comprehension assessment. However, these findings require further investigation of the online processes which contribute to successful text summarization.

2.3 Overview of methods related to investigating the L2 reading construct

2.3.1 Measuring Inferencing in Reading Comprehension

Numerous methods for assessing inferencing ability have been developed, but most measure an individual's ability to make inferences while reading a text which was constructed

around a single inference generated or written to evoke a specific set of inferences (Barth et al., 2015; Bos et al., 2016; Cain et al., 2001; Cromley & Azevedo, 2007; Singer et al., 1992; Tarchi, 2015). For example, researchers will construct a narrative which is missing a key event which is inferable from the context of the missing event. Readers would then be tasked with filling in this information is some way. Although this type of inferencing measure can be used to a great effect in identifying individual readers' inferencing ability or difficulty, these measures are less applicable to identifying where and when inferencing occurs during authentic text reading, or if inferencing contributes to successful comprehension of authentic texts. Some methods have been previously employed to understand inferencing during naturalistic reading including lexical processing measures (Potts et al., 1988), sentence processing measures (McKoon & Ratcliff, 1992), and elicitations (gap-filling targeting inferred information; Cain et al., 2001). Each of these methodologies has been employed to isolate inference generation during successful reading in a first language (L1), but this paradigm has been less utilized in L2 reading contexts.

In the L2 context, studies on inferencing ability have primarily examined lexical inferencing, or the ability to infer meanings of new words. A few studies have examined causal inferences at the text level in L2 readers. These studies have utilized short texts designed to induce inference generation, modified texts with lower and higher coherence, and self-reported inferencing strategy use to understand L2 readers' use of inferencing. Lake (2014) utilized short two-sentence texts which required an inference to maintain the coherence of the sentences. The inference either bridged the two sentences, or made a forward prediction based on the combination of the information in both sentences. Each sentence pair was followed by a true or false question which required the inferred knowledge to respond to. Lake's (2014) study found that L2 readers respond significantly faster to questions which required a bridging inference,

indicating inference making is important to L2 reading comprehension at least in terms of local coherence. However, the study did not look at inference generation during reading of longer texts. Shimizu (2009) examined bridging inferencing in a similar two-sentence coherence paradigm study. Shimizu had English learners read causally related pairs of sentences with different levels of direct causality and had them immediately recall as much as they could from the two sentences. The study found that L2 English readers with lower proficiency exhibited slower recall as the coherence of the sentence pairs required more indirect bridging. Horiba (1996) examined inferencing during the processing of larger texts, using modified high-coherence and low-coherence texts. The hypothesis is that the low-coherence texts would require more reader-responsible inferencing and would thus slow reading. However, L2 readers were not found to significantly differ in processing speed of either text type, which is the case for L1 readers. This indicates that L2 readers may utilize other compensatory mechanisms to process both high- and low-coherence texts, and that this approach does not capture L2 inferencing during reading. Feller and colleagues (2020) took a different approach to examining inferencing in multilingual readers. Their study involved surveying multilingual readers regarding self-perceptions of reading strategies. They found that higher-proficiency readers reported more activation of bridging strategies. Each of these studies measured inferencing ability using a discrete assessment or survey inference targeting inference-making ability, but no studies on L2 or multilingual readers have thus far attempted to measure inference generation as it occurred during the reading of unmodified, authentic texts, and inference generation has not been compared empirically to reading comprehension performance on tasks reflective of real-world reading assessment.

One paradigm that can be applied to inference generation during reading comprehension research involves various methods of evaluating reaction times to readers judgments of sentences. Judgements of sentences related to a previously read text, such as true/false decisions or new/old information, have been employed in various ways to examine specifically inferencing in previous research. One strand of such research involves using extended narrative texts, followed by sentences either related or unrelated to a character's goal or situation in the text (Ahmed et al., 2016; Barth et al., 2015; Graesser et al., 1994; McKoon & Ratcliff, 1992; Pike et al., 2010). The expectation is that making necessary inferences while reading the text primes the reader's response to the test sentences. This approach to measuring inferences has been useful with narrative texts and using inferencing to assess comprehension, but this methodology has not been used as frequently with expository texts or when inferencing is not the direct target of measurement. Another strand of sentence judgment tasks used to measure inferencing uses very short priming texts, only one or two sentences long, followed by a test item which is either primed by the previous text or not, but the truth of which is independent of the previous text (Ahmed et al., 2016; Graesser et al., 1994; Singer et al., 1992). Research using this approach has found that inferring causal, logical connections and activating background knowledge are part of comprehension of short passages, but this measure of inference-generation has not been applied as frequently to the comprehension longer priming texts.

2.3.2 Measuring real-time reading behavior

Understanding test takers' response behaviors and real-time cognition is critical for test validation (Borsboom, 2005) The consequential validity of tests and the decisions based on scores cannot be truly justified without knowing that the cognitive processes used to complete a test reflect the processes needed to complete a real-time task which the test qualifies one to do

(Bachman & Palmer, 1996; Bax, 2013; Khalifa & Weir, 2009). Better understanding of these processes requires non-obstructive data collection concurrent to completion of realistic assessment tasks.

The turn toward concurrent methods is ongoing in applied linguistics (Godfroid, 2019), and part of this turn is the use of eye-tracking, i.e. the collection of eye-movement behavior through simultaneous recording of readers' eyes and the object of attention. The efficacy of eye-tracking methodologies rests on the eye-mind hypothesis (Marcel A. Just & Carpenter, 1980), which assumes that "eye movements are over orienting responses that signal the alignment of attention with the object at the point of gaze" (Godfroid, 2019, p. 23). Visual attention and eye movement is strongly connected to attentional resources and cognition, and the tracking of eye movements during different cognitive activity has evolved over the years as a method to understand more about cognition, processing, and attention to language and other areas (Everling et al., 2011).

The raw information provided by eye-tracking comes in the form of *fixations* and *saccades*. While humans read, our vision not smoothly glide across a text. Instead, we move our eyes in a sequence of stops (*fixations*) and jumps (*saccades*). Fixations are any duration, longer than a pre-determined threshold (above 100 ms; Manor & Gordon, 2003), in which the eyes are relatively still. Saccades are the "jumps", or periods of active eye-movement, between one fixation and another. The position, duration, and sequence of *fixations* and *saccades* thus provide a window into the attentional processes during reading.

The granularity of these basic metrics can be refined using Areas of Interest (AOIs). By setting boundaries to certain parts of a stimulus, information about *dwells*, or *gazes*, can also be collected. A *dwell* is a sequence of fixations and saccades in an AOI, from the first saccade into

the AOI to the last saccade which leaves the AOI. Definitions of AOIs can give insight into whether or not a word is processed, the relative duration spent on certain areas of stimuli, how fixation duration modulates at different locations, and how dense fixations per dwell are on different subsets of text in a stimulus (e.g. lines and paragraphs), just to name a few metrics.

Although eye-tracking can be used to examine many different phenomena in cognition, one area where it has received extensive validation and use is in studies of L1 reading comprehension (e.g., Just et al., 1982; Rayner, 1978; Rayner et al., 1980). These studies have examined reading behavior phenomena such as relative attention to units within a text (specific paragraphs or sentences), depth of reading, jumps between fixations on words (or *saccades*), and skipped words (Jarodzka & Brand-Gruwel, 2017). These eye-tracking studies have primarily focused on lower-order reading and decoding. For lexical and syntactic processing studies, Areas of Interest are defined around specific words to understand how certain micro-textual features affect eye-movements. These rely on so-called "early measures" which include probability of fixation, time-to-first fixation, and duration of first fixation. When compared to comprehension ability, it is often found that stronger readers make fewer, shorter fixations on words than less capable readers (Ashby et al., 2005; Marcel Adam Just et al., 2018; Rayner et al., 2006). However, to analyze macro-textual processing, as one would find in reading larger portions of text (paragraphs or longer), probability of fixation and information dependent on the first fixation or gaze alone provides less information.

2.3.2.1 Eye-tracking in reading of text.

L1 research indicates that text-level reading behavior varies by task or reading purpose (Horiba et al., 1993; Kaakinen & Hyona, 2005) and by proficiency. For instance, Yeari et al. (2017) used fixation measures to examine attention to central and peripheral information

between reading goal conditions (such as reading for pleasure, reading to inform a presentation, or reading to answer questions), finding readers showed less fixation to peripheral information, relative to central information, when reading for entertainment or presentation over reading to answer comprehension questions. This indicates the importance of selective attention to specific text regions is important in at least some forms of reading assessment tasks. Jian (2017) found attention measured through eye-tracking to be significantly different between good and poor comprehenders of a passage in their L1 in various ways; notably, good comprehenders spent more time reading and integrating multiple sources of information, such as illustrations and diagrams, then poor comprehenders. Bax and Chan (2019) used eye-tracking to record reading behavior of test-takers as they completed 30 cloze and selected-response items. They found that successful readers in general made more short fixations and selectively spent more time reading relevant areas of text, whereas unsuccessful readers made fewer longer fixations in more general locations across a text. Unsuccessful readers read more slowly and focused on word level comprehension, and successful readers were more efficient when locating key information. The researchers also verified the behavior of readers with stimulated recall and survey. These results show the importance of careful reading for comprehension, as well as selective attention, especially to extratextual features, such as images. Cook and Wei (2019) surveyed the use of late measures during reading comprehension. They suggest that second-pass reading duration, i.e. rereading duration and conditional probability between areas of interest are two important sources of eye-tracking evidence of higher-order reading comprehension. However, this has not yet been applied to an L2 reading comprehension context.

2.3.2.2 L2 reading comprehension eye-tracking studies.

Despite the wide-range of eye-tracking studies focused on reading comprehension, few studies have applied eye-tracking methods to understanding reading during realistic L2 reading assessment tasks. In a rare look into how eye-movement relates to L2 reading ability, Berzak, Katz, and Levy (2018) found eye-tracking data to be useful in modeling general L2 proficiency during text reading with open-ended questions. Their reading study contextualized eye-movement in relation to overall language proficiency based on fixation on parts-of-speech and did not make further connections to comprehension of larger discourse.

Beers, Quinlan, and Harbaugh (2010) looked at rereading of students' own texts during composition and found that local and global rereadings of their own texts were predictive of text quality and writing ability, but their study looked at writing in isolation, as opposed to integrated reading-writing which one would find in summary writing, so there is still a gap in the literature regarding how eye-movements at different levels of discourse predict comprehension. Bax (2013) examined eye-movements during reading to answer fill-in-the-blank questions on the International English Language Testing System (IELTS), but found only differences in achievement at local processing levels, and used only short written production with participants of intermediate English proficiency. Prichard and Atkins looked at L2 readers eye-movement behavior in two studies (Prichard & Atkins, 2016; 2019). In their studies, they found that L2 readers of English underutilized selective reading strategies such as previewing and identifying relevant areas of text, but readers typically did use selective attention given enough time with a text. L2 readers who did apply selective attentional strategies did perform better on summary tasks. Based on these studies, it is clear the use of eye-tracking and eye-movement data to

explain process and product in reading assessment is fertile but largely unexplored territory (Conklin et al., 2018; Godfroid, 2019).

2.3.2.3 Measures derived from eye-tracking.

The information drawn from eye-tracking methods depends strongly on what measures are selected to make assertions about reading behavior. Previous research on text-processing using eye-tracking has analyzed how measurements of eye-movement behavior relate to text-level reading. Specifically, number of passes on a target, total gaze duration, and regressions are seen as important "late measures" during higher-order comprehension (Conklin et al., 2018). Mean gaze duration is also posited as an important global reading measure because it has been found to be independent of reading speed (O'Brien & de Ramirez, 2008). However, aggregate measures of eye-movement behavior may be improper conglomerations of multiple independent measurements (Orquin & Holmqvist, 2018). Total gaze duration is affected by both number and duration of fixations, so researchers using eye-tracking must be aware of which independent measurement is important to analyze.

In using eye-tracking to observe the processing of larger text, Hyönä, Lorch, and Rinck (2003) recommend not only looking at first-pass measures, but also looking at fixations measures during second-passes (i.e. rereadings or regressions). Although researchers often distinguish between the conscious process of looking back in the text ("rereading") and any saccade which jumps against the normal flow of reading ("regression"), researchers agree either that the need to reread motivates regressions or that the natural process of regressing motivates rereading (Booth & Weger, 2013). Rereading in this study is a specific type of regression between macro-textual features plus any forward saccades following a regression but not ahead of the initial regression site. Thus, the term rereading is used for this behavior rather than simply regression to capture

the meaning-building nature of the process. Examples of rereading are when the bulk of careful paragraph reading is done during second pass, after skimming or scanning, or when a reader jumps back to a previous paragraph soon after beginning a new one to resolve or complete comprehension. Jarodzka & Brand-Gruwel (2017), in an extensive review of eye-tracking and reading behavior, explain that text reading differs from local text parsing in being more careful, with more attention to each word, less skipped words, and shorter saccades between fixations. Text level reading also involves fixation on meta-textual objects such as pictures or diagrams which may be integrated with textual information during reading.

2.4 Expected findings

With regards to research question 1a, previous research using the sentence verification task paradigm, requiring true or false responses from participants, have typically found that stimuli which are more congruent with earlier stimuli are *primed* by the earlier stimuli, and are thus responded to with greater ease (Collins & Quillian, 1970; Knoeferle et al., 2011; Macleod et al., 1978; Ratcliff & McKoon, 1978). The priming stimuli in previous research were typically not much longer than the target stimuli, with word-to-word or sentence-to-sentence priming found when stimuli shared certain qualities, but stimuli priming sentence judgment times could also be pictures (Clark & Chase, 1972). However, in the current study, the sentence verification task involves a series of sentences, half of which were primed by, but not copied from, a reading passage, and half of which were control sentences. The participants had to decide if the sentences were true or false. Although different from the typical sentence verification task, it can still be hypothesized that related sentences will be responded to more quickly than unrelated sentences.

For research question 1b, it can be hypothesized that inference generation, as measured by relatively faster reaction times to inferred sentences during a post-hoc sentence verification

task after reading a priming text, will relate positively to score. If the reading comprehension measures (MC tasks, cloze tasks, and summary writing tasks) tap into higher-order comprehension processes and push readers to create a mental model of the text they are reading, then readers will naturally make certain inferences as part of successful comprehension. Thus, it is hypothesized that there will be correlations between faster reaction times to inferable sentences and higher scores on the comprehension tasks, and further that inference reaction speed will be a significant predictor of score in linear modeling. These effects are hypothesized to be stronger in the summary task, which more explicitly pushes participants to perform text modeling processes for successful task completion.

For research question 2a, significant differences between eye-tracking measures are expected between the three tasks. The cloze task, due to local constraints on each blank to be filled, is likely to elicit careful local reading, and longer fixation times and denser fixations per dwell are expected. The MC task is hypothesized to elicit expedient, linear reading, in the form of more fixations per dwell in each line, less fixations per word overall and re-reading, and less global metrics such as length of saccade and transfer between text and task. This hypothesis is considered because of the selective goal-setting provided to the reader by the questions, so global careful reading may not be necessary. For the summary task, it is hypothesized that reading will be more global and careful, as text modeling is more critical to completing the task, and that this will manifest in higher fixations per word overall, more transitions between text and task, and longer saccades as readers make connections across distant parts of texts.

Little research has compared reading comprehension score outcomes with eye-tracking, so hypotheses regarding how eye movement will affect score are not as clear. It can be hypothesized that more better readers are more efficient readers (Grabe, 2009), so shorter

fixation duration will likely predict higher scores in each task. However, task specific eye movements may also become relevant predictors.

There are clear avenues for further exploration of the intersections between real-time reading behavior, L2 proficiency, L2 reading assessment formats and reading performance. While the possibility exists for more complex modeling comprehension scores by utilizing L2 proficiency groups and interaction effects between the measures described, the above lines of inquiry are fairly exploratory in nature and further research questions outside those covered by this dissertation are considered in the conclusion chapter. The operationalization of the constructs in these questions, the data collection procedures, and the data analysis are outlined in the methods chapter (chapter 3). The following section contains further literature review, going deeper into the background of the reading comprehension construct and L2 reading assessment.

3 METHODS

3.1 Research Design

3.1.1 Participants

A total of 102 (68 female) students were recruited from tertiary education programs at a large Southeastern United States public university. The sample size was derived from an *a priori* power analysis, calculated with G-Power, indicating that at least 98 participants were needed to reach 80% power in a linear model with 7 predictor variables, given an alpha value of .01 and observed effect size of 0.2. A further consideration was related to the different reading comprehension test forms employed in the study (explained below), of which there are 18 (3 tasks x 6 topics), and so a number divisible by 18 is necessary to balance across test forms. Thus 102 participants were recruited.

Participants were screened for inclusion based on self-reporting experiences with formal English language education, either within or outside of the U.S., and reporting no cognitive disabilities which may interfere with their reading ability in any language. This inclusion criteria was used to ensure recruitment of a diverse population similar to students who have been successful in standardized English proficiency tests such as the TOEFL and IELTS, on which reading is a component. Unlike TOEFL and IELTS test-taker populations, the students in this study were all matriculated at the time of participation. Non-matriculated students were not selected because they were not available at the time of this study. This means that the sample in this study is more reflective of the successful test-taker population rather than a general test-taker population. As such, various individual abilities were measured for each participant, including a L2 morpho-syntactic proficiency test, to gradate participants beyond what is implied by their already sufficient proficiency test scores.

Participants were either international students or multilingual English speakers with a history of formal English educational background. Their ages ranged from 19 to 52. 43 students were in undergraduate programs, 55 were in graduate programs, and 6 were in an intensive English program. They represented 29 language backgrounds, with the most common being Mandarin (n = 21), Spanish (n = 17), Korean (n = 9), Telugu (n =8), Cantonese (n = 6), Urdu (n = 4), Vietnamese, (n =4), and 21 other languages with three or fewer participant representatives (n = 27). Participants had spent on average 4.67 years in English speaking countries and had taken an average of 5.1 years of formal English classes.

3.1.2 Selection of texts

The experimental test procedure involved reading introductory academic texts taken from various fields of science (applied sciences, natural sciences, social sciences), akin to what one would find in a textbook introduction for an introductory class to an academic subject. Texts were selected from free textbook resources available from Georgia Virtual Learning (http://www.gavirtuallearning.org/Resources.aspx, n.d.), which provides online textbooks for high school students in the U.S. state of Georgia. Six texts were selected, and a form was written of each task type (multiple-choice questions, cloze, and summary) for each text. The two applied sciences texts were “Biotechnology,” which was about the application of DNA research to medicine and other fields, and “Microscope,” which was about the development, functions, and applications of the compound microscope. The two natural sciences texts were “Water,” which was about the chemical properties and importance of water on Earth, and “Hunger,” which was about the biological, psychological, and cultural motivations of feelings of hunger. The two social sciences texts were “Choices,” which was about the economic principles of trade-offs and opportunity cost and their use in decision-making, and “Attitudes,” which was about cognitive

dissonance and the way our roles and actions influence our attitudes and beliefs. Each participant was shown one text from each category, and thus read three full texts. Each text was presented in one task form, so participants completed one of each task forms. See Appendix D for the texts in each format.

Texts were not modified, although they were taken from longer contexts. Texts ranged from 315 to 350 words, consisted of four paragraphs, and were selected based on their content, intended level (grade 11), and lexical and syntactic complexity. This grade level for texts was chosen for multiple reasons. The availability of open-source, level-comparable textbooks from which passages can be drawn is higher for high-school textbooks than college level ones. Also, although all participants in this study are at the university level or above, a priori knowledge of the participant sample's reading level was unattainable, so high school-level texts were chosen to more carefully ensure approachability of the texts to the participants. The selection of high school level texts also adds the benefit for easier comparability, since the reading level for the texts was measured by Flesch-Kincaid Grade Level readability index, scores from which are less precise at college reading levels. Mainly, texts below the university reading level were selected to increase the expected ability of the participants. While these texts were below the expected reading demands of the participants, they should allow participants to devote mental resources to higher-order comprehension. Choosing texts at a higher difficulty level while still utilizing authentic, unmodified texts from specific topic domains may have required too much lexical inferencing, i.e. the guessing of unknown words, and thus reading may have been too reliant on the background and specific vocabulary knowledge of the readers.

Texts were further analyzed for lexical sophistication and discourse complexity using the Natural Language Processing tool TAALES (Kyle, Crossley, & Berger, 2018) to ensure that

texts were similar in terms of vocabulary demands. Specific lexical sophistication indices related to text level were analyzed, including average word concreteness, average age of acquisition of words, and range and frequency in the academic subcorpus of the Corpus of Contemporary American English (Davies, 2008). These metrics are calculated as the average per word for each text. For more details regarding these metrics, see Kyle, Crossley, and Berger (2018). The averages for each metric were calculated, and a one-sample t-test was used for each metric to ensure that the lexical sophistication of the texts was within a homogeneous range. Insignificant p-values show texts are not significantly different from the average for a given metric, where significant p-values would show that at least one text is different from the others for that metric. Importantly, no text was significantly deviant from the mean for any of the metrics, including intended reading level. These comparisons are presented in Table 3.1.1.

Table 3.1.1 Text reading level and lexical sophistication

Measure	M	SD	*t*	*p*
Concreteness (Brysbaert)	2.910	0.187	-0.497	0.680
Age of Acquisition (Kuperman)	7.053	0.544	-0.190	0.572
COCA academic range	0.318	0.049	0.747	0.244
COCA academic frequency	883.441	233.710	-1.018	0.822
Flesch-Kincaid	11.229	1.225	-0.816	0.774

3.1.3 Selection of Reading Comprehension tasks

The primary reading comprehension tasks involved reading an academic text (described below) and completing reading comprehension items. Participants completed three readings, and each reading text was accompanied or augmented by a different comprehension task: MC questions, a cloze task, or a summary task. These tasks were chosen for their prevalence as

language assessment formats, their different degrees of response constraint (with MC items being fully constrained and cloze and summary formats being less constrained), and discreteness (with MC and cloze forms having discrete-point scoring and summary requiring rubric-based ratings).

Five-item multiple-choice (MC) test (one for each of six topics) was developed by the researcher and a group of linguists trained in assessment design. Specifications were provided for how multiple-choice questions should be written based on Day & Park's (2005) taxonomy of reading comprehension items. The specifications entailed writing five items for each text to target multiple aspects of comprehension and limit the targeting of language features. The questions included one question addressing the passage's main idea, two questions addressing specific details, one asking participants to make an inference connecting pieces of information in the passage (bridging inference), and one which asked readers to make a prediction or elaborate outside of the literal information in the text (elaborative inference). The specifications also ensured that each multiple-choice question included three options with only one correct option. Multiple-choice questions typically have either three or four potential options, with three being the optimal number of options for reliability and discrimination (Loudon & Macias-Muñoz, 2018; Rodriguez, 2005). For each question the incorrect answers were designed in a specific way such that distractors attracted different types of poor comprehenders (Carlson et al., 2014). One distractor was an attractive answer to readers who read expeditiously and over-rely on their own assumptions, and one distractor would relate to linguist cues in the text and would be attractive to readers who read slowly and carefully but perhaps did not capture propositional meaning while reading. Thus, each question had three answer options: one correct option, a distractor targeting irrelevant text information, and a logically plausible distractor with little logical linkage

to the source text. Pilot testing was conducted at an Intensive English Program at a large U.S. university. From the piloting data items which were too difficult or too easy were modified based on the piloting.

For the cloze tasks, participants were presented with a clozed version of an academic text with 15 words replaced by gaps. Participants were instructed to type a word into each blank which maintained the coherence of the passage. Although cloze tasks often involve systematic or random deletion of words in the text (Carrell, Carson, & Zhe, 1993), this procedure limits the validity of cloze tests as assessments of reading comprehension. Rational cloze tests, where specifically content words and coherence-maintaining words are deleted are preferred when directly addressing reading comprehension macroprocesses (Greene, 2001; Kleijn, 2018). Thus, the cloze words were selected with this in mind, targeting content words related to the text topic (but the absence of which would not eliminate coherence of the text) or words which create coherence links in the text, such as connectives and repeated words. During reading, participants typed words into highlighted blanks which they believe to best complete the text. They navigated between blanks using direction buttons. Rating procedures are discussed in the data analysis section below.

For the summary tasks, participants were presented with an academic text and given a textbox to the right of the text into which they directly typed their answer. There are multiple types of summary writing, and, to make this summary task more grounded in academic expectations, the exact summary task is similar to the ‘brief account’ summary format detailed by Seidlhofer (1990). This type of summary is not a mere linguistic reduction or truncation of a source text, but instead a purposeful yet brief transmission of text information to a secondary audience. Participants were instructed to write summaries directed at a hypothetical fellow

student taking a course on the same topic as the text, and to keep the summary between 100 and 150 words. Providing a hypothetical audience is intended to push the participant to write a summary around the necessary content of the given text (Seidlhofer, 1990), and not necessarily its verbatim linguistic features (i.e., recall). This also gives the summary task an explanatory function for the reader, which can be part of successful comprehension and text learning (McNamara, 2004). Summary rating is discussed in the data analysis section below. Appendix D presents each text used in full, as used in each format. MC questions can also be seen there, and the summary prompt can also be found.

3.1.4 Operationalization of higher-order skills

3.1.4.1 Inferencing

To measure inference generation, a sentence verification task was administered after each reading comprehension task was completed. The current study used a novel approach to sentence judgment tasks which synthesizes previous methods described in chapter 2. The task used in this study involved sentences which were either primed (related) or not primed by the text reading comprehension task the participant had completed. The test sentences were either true or false, and the veracity of the sentences was determinable without having read the priming text, although having read the text would help in this determination. In other words, the sentences are general enough to comprehend without reading the text, but the topically related sentences represent information critical to comprehension of the text. Similar to previous narrative-focused studies, this method compares long text primes to related and unrelated information, but similar to previous short text inference studies, this method uses test sentences which have real-world truth values independent of, but related to, the priming text. In this way, the method can measure

whether or not inferences based on real-world knowledge and logic are activated during the reading of extended expository texts.

An example of this procedure is presented in Figure 1. After a participant finished a reading a text and completing the concurrent comprehension task, the text's respective sentence verification task began on a new screen and involved reading a series of 16 sentences. In this way, the influence of completing the reading comprehension task can be may be drawn from reaction times to sentences presented to the participants after the reading text. For each sentence, subjects indicated whether the sentence was true or false. Knowing the veracity of the sentences was not contingent upon understanding of the texts, and the truth values of the sentences were rooted in real world facts or falsities. Eight sentences were by primed by the text, and eight were irrelevant control sentences. The true/false and related/unrelated categories overlapped, creating a matrix of four sentence conditions: true-related (inferences), false-related (inversions of inferences), true-unrelated (true control sentences), and false-unrelated (false control sentences).

The 16 sentences for a given text were presented to participants in a random order. None of the sentences appeared verbatim in the texts, MC questions, or task prompts, nor were they a paraphrase of any specific proposition in the texts, MC questions, or task prompts. Instead, the true related sentences represented ideas which would positively contribute to modeling the text if inferred during reading (e.g. in the text participants read on "Biotechnology", the sentence "Every living thing contains unique genetic information."). The false related sentences (e.g. "Genes change naturally throughout an average person's life."), may slow comprehension time with respect to true related sentences, but if responded to correctly, should still be responded to faster than unrelated false sentences. The unrelated sentences were also true (e.g. "Scientific procedures require precise and accurate data.") or false (e.g. "Light and sound waves never

change direction after hitting an object."), but the information within would not be necessary to comprehend the text. The unrelated sentences for one text instead came from another text's related true and false sentences.

Each sentence was between seven and thirteen words (45 to 69 letter characters), following McKoon & Ratcliff's (1992) task. When subjects indicated their response for a sentence, using a button press, there was a 1000 ms pause, followed by a prompt to press a key to see the next sentence. After a key was pressed, a screen with non-language characters appeared for 1000 ms to have participants re-fixate on the center of the screen, and then the next trial sentence would appear. Backward masking, i.e. covering the target stimulus with non-target a stimulus to force processing within a fixed time frame, was not employed since the window of time for masking to be effective (30ms; Breitmeyer & Ogmen, 2007) is too narrow for sentence verification. Further, distractor tasks between verification sentences was not employed as it risks cognitively isolating verification sentences from the reading text by extending the already high stimulus-onset asynchrony of the sentences (Harley, 2008, p. 171). The trial procedure is outlined in Figure 3.1.1 and Appendix E presents information about the 16 sentences for veracity judgments corresponding to a reading task.

3.1.4.2 Text-level reading eye-movement behavior

In addition to the use of post-hoc measures, the examination of real-time reading behavior requires the collection of data concurrent to the activation of reading processes. Reading behavior was recorded using eye-tracking methodology, and eye-movement behaviors during reading were operationalized using metrics gathered via eye-tracking. During text reading, participants were seated at a computer about 2 feet from the computer screen and completed three reading comprehension tasks.

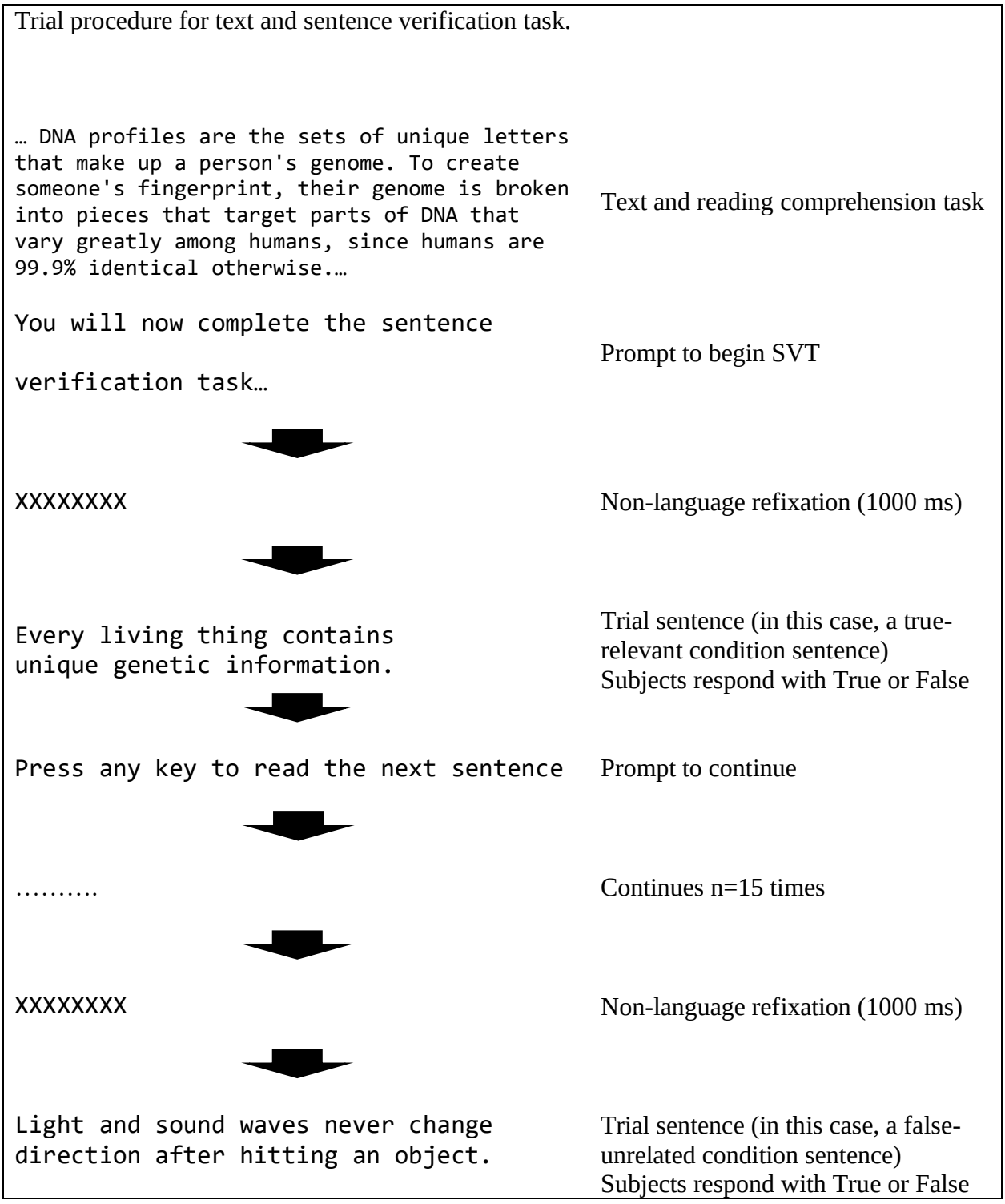


Figure 3.1.1 Experimental trial sequence for the sentence verification task.

ASL EyeTrac 6 software automatically gathered raw data from participants' location and duration of fixations, and the length of *saccades* (or "jumps") between fixations in each reading task. Fixations are the momentary stops made during reading, and they are the basic metric of attention in tracking eye-movement. For this study, any pause in eye-movement greater than 100 ms (.1 seconds) was regarded as a fixation (Manor & Gordon, 2003). Post-hoc areas of interest (AOIs) were designated around each paragraph and line of a text. Although lines are not linguistically relevant portions of text, they are the longest span of text for which a linear sequence of fixations can occur before readers are forced to make a *return-sweep* (a saccade which brings a reader back across the text to the start of a line), so fixations per line dwell gives a sense of how many fixations on average occur in unbroken, linear line reading. AOIs specific to each task were also constructed on multiple questions, on Cloze blanks, and on the text box for summary-writing.

Using raw data from the location and duration of fixations, many different often-used eye-tracking metrics can be calculated including probability of fixation, single fixation duration, and time-to-first fixation. However, these measures are associated with "early" processing of local contexts (words and short sentences) rather than "later" processing associated with integrating large portions of text as found in this study (Cook & Wei, 2019). Later processing of texts, involving cognition beyond lexical recognition and phrase/sentence parsing, involves global eye-tracking measures suited to understanding text-level stimuli. Time to first fixation and probability of fixation may be valuable for understanding eye-movement behavior properties of certain lexical items, but they have less value in understanding larger portions of text (e.g. the probability of a participant looking at the first paragraph in any of the reading tasks was 100% for this study). The measures calculated with eye-tracking for this study are thus measures which

have previously been associated with text-level reading comprehension: mean length of saccade, total number of fixations per word across a trial, total time spent rereading portions of text. Mean length of saccade was calculated as the average Euclidean distance between each sequential pair of fixations. Fixations per word was calculated for text areas of interest as the total number of fixations made on the reading passage divided by the number of words in the passage.

The calculation for rereading was more complex. Rereading is not marked simply by the second gaze on an area of interest, as participants often began the tasks by making sporadic fixations across the screen before settling on a place to beginning the task, be it the instructions, beginning of the text, or elsewhere. These fixations were thus removed from calculation as "first-passes" into any areas of interest. Scan-path videos were used to manually determine when a participant's first pass into line and paragraph areas of interest was after reading had begun. Rereading duration for each area was calculated as the total time of gaze in an area of interest excluding the first pass and any stray fixations before the first pass. Although lines of text are not a true structural unit of texts, a measure of the rereading which took place within paragraphs was necessary, and areas of interest around sentences were not geometrically consistent enough to manually draw areas of interest in the eye-tracking results interface.

To examine if global and careful text reading occurs, fixations per word for the entire text-task trial were collected, as well as mean length of saccade per trial. Within each AOI, data collected were average number of fixations per pass through the AOI, mean fixation duration, and number of fixations during rereading for each text, each normalized for number of words per AOI. Rereadings for paragraph AOIs were measured by calculating gaze duration in AOIs excluding the first pass dwell.

Next, since reading text in this study is done alongside completion of comprehension tasks, we are also interested in how attention is given to task areas, so total number of fixations per word in the task areas and number of transitions between gazing at text and gazing at tasks (henceforth *transitions*) were also calculated. Fixations per word was calculated for task areas as the total number of fixations made on the task areas divided by the number of words in the task area. For cloze tasks, the number of gaps in the task was always 15, and each was filled by a single word, so the number of words per task was considered to be 15. For MC tasks, the number of words was counted for question stems and all answer choices. For summary tasks, the individual summary word lengths were used as number of words, which also controlled for the length of summary. This is an admittedly rough approach to measuring for task length in words, but it adequately reflects the size of difference task areas to make some comparisons across tasks possible. Transitions were calculated as the number of times a participant shifted their gaze from the text area to the task area. For the Cloze task, this meant moving from a word in the text to the one of the gaps where a word was to be entered. For the MC task, this meant shifting gaze between the text and the question area. For the summary task, this meant shifting gaze between the text and the summary area.

Lastly, the type of reading associated with comprehending and learning from text is considered *careful reading* by Urquhart & Weir (2014), which is slower and more linear than *expedient reading*. Thus, eye-tracking measures which may relate to careful reading are also calculated. These include average text fixation duration, average task fixation duration, average number of fixations per dwell in line reading, and average number of fixations per dwell in paragraph reading. Average fixation duration is included as longer fixation durations can be an indicator of careful, slow reading (J. Wang et al., 2018). Fixations per dwell is a measure of how

many fixations are made in a particular area of interest between entering the area and leaving the area via saccade, with greater fixations per dwell indicating more careful attention to the area of interest (Holmqvist et al., 2011). Mean fixations per dwell by line was calculated as the average number of fixations between beginning a dwell in a line area of interest before shifting gaze away from the line. Paragraphs are a more valid feature of discourse, and fixations per dwell provide a measure of the level of careful, reading at the paragraph level, although line rereadings may occur in within one paragraph dwell. Paragraph dwells may contain multiple line dwells. Mean fixations per dwell by paragraph were calculated as the average number of fixations between beginning a dwell in a paragraph area of interest before shifting gaze away from the paragraph.

It is important to note that although certain eye-tracking measurements may be related to certain underlying constructs, it is not the intention here to causally equate, before the fact, eye movements with underlying behavior. For example, although length of saccades is associated with global attention rather than local attention, it could also indicate distraction and lack of attention depending on the direction of saccades as well. Additionally, measures such as mean fixation duration may be tapping into underlying individual differences rather than conscious effort to read more carefully. This is mitigated somewhat by the within-participants comparisons that make up a portion of eye-tracking analyses in this study.

Before recording, the eye-tracking camera was calibrated to the individual participant. Accuracy to within .2 inches was ensured before recording began. If a participant fell out of calibration during the procedure, the researcher could make small adjustments to fix the recording, or else pause the experiment to reorient the participant. In addition to calibration, visual scan-path data was also collected. This visual representation of the path of fixations made

by participants was used to ensure that the recording was aligned with the image presented on the screen. Data which included too many unexpected fixations (focused off-screen or on blank space) or which appeared to be skewed with respect to the orientation of the text presented on-screen were discarded.

3.1.5 Individual factors of reading ability

3.1.5.1 Morpho-syntactic Proficiency and Vocabulary

Academic reading ability depends heavily on general language proficiency and vocabulary size, which is developed alongside academic literacy with many multilingual students and scholars (Laufer & Nation, 1999). English proficiency would have ideally been gathered using reported performance on a proficiency test such as IELTS or TOEFL, but due to the diversity of the participant sample, not all participants had a recent comparable proficiency score or had access to their score. Instead, a brief 18-item gap-fill test targeting morpho-grammatical knowledge and vocabulary size was administered as a language proficiency test. This test involved deleting the second half of target words in otherwise coherent sentences to create a gap-fill task. The test is based on the productive orthographic vocabulary size tests (Laufer & Nation, 1999) which have been found to strongly predict reading comprehension in a second language (Cheng & Matthews, 2018). Specifically words from the 6000 to 8000 most frequent words in COCA Academic (Davies, 2008) needed for academic reading at the university level (Crossley et al., 2016; Kyle & Crossley, 2015), were targeted in a gap-fill task with a set of 18 sentences which contain target words (Appendix B). The words which were targeted involved a range of inflectional and derivational morphological endings to also tap into grammatical knowledge in addition to vocabulary size. Scoring was done using an answer key. Correct answers were marked for 1 point, and answers which did not maintain the intended meaning were marked as 0

points. Item left blank were marked as 0. Answers which matched the key semantically but had incorrect inflection or part-of-speech marking were given a half point (.5). Each participant received a score out of 18. Reliability statistics were calculated for the morphosyntactic proficiency test in the following chapter.

3.1.5.2 Reading and Typing Speed

Reading fluency is an important lower-order literacy skill (Gauvin & Hulstijn, 2010; W. Grabe, 2009; Stoller et al., 2013), and should be measured and controlled for in any study of higher-order reading processes. Additionally, reading fluency has been found to exhibit effects on eye-tracking measures in monolingual data (Taylor & Perfetti, 2016). Reading fluency was thus measured by words per minute read during a silent reading of a 375-word 12th grade-level academic text about volcanoes (not one of the texts included in the main procedure). This text was followed by four comprehension questions just to ensure the participants read intentionally; however, this was not figured into calculations as a measure of comprehension.

Although the way in which tasks are scored is intended to mitigate the influence of productive skills, production fluency remains connected to comprehension through the broader construct of literacy (Belcher & Hirvela, 2001). Due to the productive aspect of the cloze and summary tasks, a measure of L2 writing ability is warranted, but was impractical given the time demands placed on the participants. In lieu of a comprehensive measure of L2 writing proficiency, typing speed was gathered as a measure of production fluency. The fluency with which participants produce responses may also affect their performance (Barkaoui, 2014). As such, a measure of typing speed was included as a baseline individual difference. Participants were asked to type as many words as possible in 60 seconds. The words to type were randomly selected words which appeared on the computer screen. Participants were given real-time

feedback as they typed regarding whether a word was typed correctly and when the word was completed. The typing speed test was taken from the free typing speed test at livechat.com (LiveChat, 2016).

3.1.5.3 Reading Motivation

Because this study focuses on reading comprehension tasks as purposeful, an important factor in measuring academic reading comprehension is motivation for reading. Motivation for reading has been found to contribute to reading comprehension skills in previous research (Schaffner & Schiefele, 2013; Wigfield & Guthrie, 1997). A reading motivation survey was administered to participants before the reading trials and consisted of a brief discrete-point item survey, using a 5-point Likert scale, regarding reading motivation. This survey, developed for this study, included 10 items measuring intrinsic and extrinsic reading motivation (five each). Intrinsic reading motivation refers to internal or personal reasons for reading where reading is a means to its own end (enjoyment, personal enrichment), and extrinsic reading motivation refers to external or practical reasons for reading where reading is a means to some other end (career-usefulness of reading, social engagement through reading). The items were subjective agreement items (e.g. “I enjoy reading about topics which I have discussed with others.”). These items were derived from previous surveys of motivation (Ryan & Deci, 2000; Wigfield & Guthrie, 1997), with some reductions made for the sake of practicality. The survey was validated using confirmatory factor analysis to ensure that the items targeting the different motivational constructs factored together. These results are presented in the following chapter. The survey instrument is presented in Appendix C.

3.1.5.4 Reasoning

Inductive reasoning ability has been found to predict reading comprehension in previous research (Klauer & Phye, 2008). Specifically, inductive reasoning refers to one's ability to extrapolate information beyond what is presented, and to notice patterns and regularities. This ability to draw conclusions from observations has been shown to be related to inference-generation skill (Schaffner & Schiefele, 2013). For this study, inductive reasoning was measured using an incomplete series test (123test, n.d.), where the first three items of a patterned sequence of shapes were presented, and participants filled in the fourth item in the sequence from four options (See Figure 3.1.2 for an example). The test consisted of ten dichotomously scored items. Reliability for the test was calculated using Cronbach's Alpha, and these results are presented in the following chapter. The test took approximately five minutes to complete and was administered via computer using a web browser. As the focus of the current study is on the contribution of inference making to reading comprehension, inductive reasoning scores were included as a control variable in models of comprehension score.

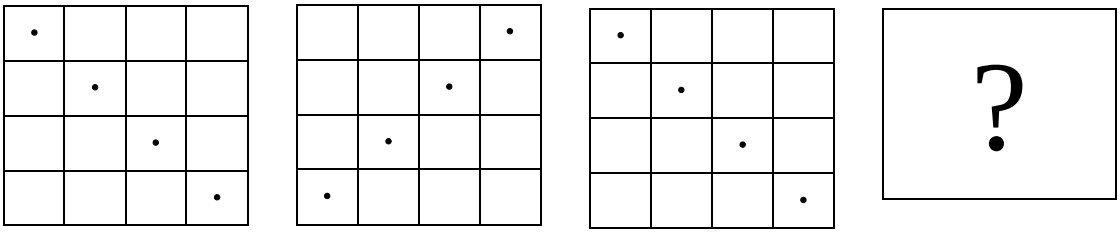


Figure 3.1.2 Example of an incomplete series test item using dot patterns in matrices.

Note: The intended answer in this case is identical to the second member of the series.

3.2.5.5 Working Memory

Working memory capacity is an important cognitive ability to measure in the current study because it has been found to contribute to reading comprehension and inference-making ability in monolingual readers (Cain et al., 2001; Calvo, 2005; Carretti et al., 2009) and multilingual readers (Alptekin & Erçetin, 2010; Erçetin & Alptekin, 2013; Joh & Plakans, 2017; Lipka & Siegel, 2012). Since the SVT for inferencing was administered after text reading, working memory capacity is a potential moderator when comparing SVT data to comprehension data. Working memory was measured using a 2-back test. Although n-back tests are contentious as measures of working memory capacity (Jaeggi et al, 2010), there is evidence to suggest that they work as a measure of working memory in adults (Haaveit et al, 2010; Tsai, 2014) and visual memory capacity (Gajewski et al, 2018), which is appropriate for the current study focused on reading.

In the 2-back test, participants were shown a series of simple images. At each image, participants compared the current image to the image they saw two images previously. They indicated through mouse click whenever the current image matched the image shown two images previously. They saw a total of 35 images, with each image presented for one second. 15 2-back matches were randomly distributed in the sequence of pictures. Scores were reported as a percentage of correct responses to total images shown minus 2. Reliability is presented in the following chapter.

3.1.6 Data Collection Procedure.

The procedure consisted of two main components. First, after meeting the researcher and signing informed consent, participants provided demographic information (i.e., age, academic level, language background; see Appendix A) and completed the individual difference tasks

including reading motivation, L2 English proficiency, reading speed, logical reasoning, and working memory. Each test was selected or designed with a five-minute time limit in mind.

Second, after the individual difference measures were completed, calibration for the experimental trials began. Participants were seated in front of a 22-inch monitor at a distance of about 24 inches. The participants rested their chin on an adjustable chin rest and made any adjustments to their seating needed before continuing. Participants entered all responses and navigated the stimuli using keyboard buttons, and the mouse was disabled during the trials. The eye-tracker was calibrated by having the participants look at fixed calibration points on a gray background while the eye-tracking camera recorded their gaze. Data from nine fixation points were used, and the researcher confirmed that the calibration was successful by asking the participant to look at specific points on the screen and verifying the camera's accuracy within .25 of an inch.

The experimental trials involved reading three texts and completing one of three possible reading comprehension tasks for each text: a multiple-choice reading task, a cloze task, and a source-based summary writing task. During each reading and comprehension task, eye-movement data was gathered with an eye-tracking camera. The camera recording eye movements was an EyeTrac 6 eye-tracking system from Applied Science Laboratories, which measures eye movement with a 60Hz sampling rate. Participants used both eyes in the study, but only measurements from the dominant eye was taken.

The experimental stimuli were presented in Paradigm stimulus presentation software. This software allowed participants the capability to self-pace as they progressed through the trials and use multiple response devices simultaneously (keyboard and response box), and it offered the means to practically present text and task-response on a single screen which would

not scroll or change position. The presentation began with an introduction section explaining the sequence of tasks and instructions for each reading task. The researcher read this aloud along with the participant. After the participant confirmed they understood the instructions, they pressed a key to move onto a training section. In the training section, participants completed short versions of each task, including reading a one-paragraph-long text with two blanks to fill (to practice the cloze procedure), reading a one-paragraph-long text and answering two multiple choice questions, reading a one-paragraph-long text and writing a short summary, and reading 4 sentences (one on the screen at a time) and respond true or false. This way, the participants were familiar with each type of comprehension task as well as the sentence verification task.

After the practice section, participants moved on to the first task. For each text and task combination, the instructions for the task were presented, and then the text would appear. Texts were presented to participants in full, statically on the screen (i.e., texts fit on the page without requiring a scroll bar to navigate the text) in double-spaced, size 14 Consolas font (a fixed-width font). The text occupied roughly the leftmost 70% of the screen, with a one-inch margin, and the right-most 30% contained either comprehension questions available at the start of reading in the multiple-choice trials, a text box to enter a summary in the summary trials, or was blank in the cloze trials aside from the instructions. Text reading and comprehension tasks were completed simultaneously in each case. After each comprehension task, the SVT related to the text was presented, and participants advanced through each sentence one screen at a time, responding true or false. This is completed after the entire related reading comprehension task trial to ascertain the influence of completing the reading comprehension task on the reaction times in the SVT. After participants responded to the 16 sentences, they moved on to the next text.

Table 3.1.2 presents an overview of the data collection procedure. A counter-balanced design was set up so that each participant read three different texts, one in each task format, with the six texts being used evenly across participants and tasks. The order in which tasks are completed was also counterbalanced. Table 3.1.3 displays the ordering of task and topic presentation to participants. For example, participant 1 would first complete the multiple-choice form for the text "Biotechnology", then the cloze form for "Water", and lastly the summary form for "Attitudes". In this way, each of the eighteen task-topic combinations (3 tasks x 6 topics) was seen an equal number of times across participants. Each reading task and its respective sentence task are intended to take approximately 25 minutes to complete, meaning a total of one hour and 15 minutes for reading and sentence verification naming.

Table 3.1.2 Data collection sequence.

1.	Demographic survey
2.	Reading motivation survey
3.	English proficiency measure
4.	Reading fluency measure
5.	Logical reasoning measure
6.	Working memory measure
7.	Reading and comprehension task 1
8.	Text 1 verification task
9.	Reading and comprehension task 2
10.	Text 2 verification task
11.	Reading and comprehension task 3
12.	Text 3 verification task

Table 3.1.3 Task and text topic order.

ID	Task Order			Topic Order		
1	MC	Cloze	Summary	"Biotechnology"	"Water"	"Attitudes"
2	Cloze	Summary	MC	"Choices"	"Microscope"	"Hunger"
3	Summary	MC	Cloze	"Water"	"Attitudes"	"Biotechnology"
4	MC	Cloze	Summary	"Microscope"	"Hunger"	"Choices"
5	Cloze	Summary	MC	"Attitudes"	"Biotechnology"	"Water"
6	Summary	MC	Cloze	"Hunger"	"Choices"	"Microscope"

7	…	…	…	…	…	…

3.1.7 Scoring

Each participant's responses for each task were scored in an appropriate manner. MC task responses were scored automatically. Cloze tests were each scored by a trained rater and the researcher with an answer key using an acceptable response scoring method. Each cloze blank had a known, intended response based on the source text, but near-synonyms of various degrees of specificity were allowed for full or partial credit, and words that fit the context semantically but were grammatically incorrect were also worth half a point. Although the cloze tests were scored discretely with each of the 15 blanks counting as 0, .5, or 1 point(s), human raters were chosen rather than automated rating due to the occurrence of acceptable synonyms and correct words with the wrong form which were worth partial credit. Raters had a chance to decide if a non-keyed response still created a coherent text segment. The researcher and each rater conferred about non-keyed response scores to reach agreement on scoring. Each correct response to a blank in the passage was to be given a point, for a maximum score of 15 points.

Summary rating was performed by trained raters. Raters were all graduate students in an applied linguistics department, and raters were compensated for their rating. Summaries were rated using an analytic rubric developed by the researcher (see Appendix F for the full summary rating guidelines) and informed by Taylor (2013). Although Taylor (2013) used a holistic rubric to rate gap-filling summary tasks, this study uses an analytic rubric based on constructs used in Taylor's rubric. This rubric is used to measure summary quality on the constructs of content accuracy (whether or not a summary was accurate and complete with respect to the source text),

level of modeling (how well the summary distinguished between main and subordinate ideas, and generalizes across smaller details), task completion (to what degree the summary fit the word length parameters, was organized with respect to the source text, and conveyed useful and coherent information to a hypothetical peer), and language quality (including linguistic accuracy and use of source text). Only accuracy, modeling, and task completion are used as measures of comprehension (language is used to control for productive language ability). The language score component was only included on the rubric to mitigate the effect of raters' judgments of productive language quality on their assessment of the reading comprehension components and was not intended to reflect overall comprehension score.

Each summary was given a separate score on a scale from 0 to 4 for each construct, and each summary was rated by at least two raters. In the case that ratings from the first two raters differed in any category by more than one point, a third rater provided a third rating for the summary. The average of the closest two ratings for a given rubric construct were used as the final score, and an additional Total Comprehension score was calculated as the sum of the accuracy, modeling, and task completion ratings for each summary. Scores were analyzed for inter-rater reliability using Cohen's Kappa, and additionally analyzed for rater fit and rubric reliability using Multi-faceted Rasch Analysis (Linacre, 2002).

3.2 Analyses

To address each of the research questions described in the previous section, a series of statistical analyses were performed.

3.2.1 Research question 1

3.2.1.1 Do examinees respond significantly faster to sentences inferable from a text than to unrelated sentences after reading the text and is this mediated by reading comprehension tasks?

To answer the first part of research question 1, of whether inferable sentences are primed by the text reading, reaction times to items with correct responses during the sentence verification tasks were gathered and controlled for length of sentence. These reaction times are modeled as the dependent variable using Linear Mixed Effects (LME) modelling, with sentence type as the single fixed effect and subject as the random effect, as subjects gave multiple responses for each independent variable category. Sentence truth value (true or false) and text-relatedness (related or unrelated) were categories for the fixed effect.

If correct responses to true/false related sentences have significantly faster reaction times than true/false unrelated sentences, this provides evidence that generating inferences was a component of L2 expository text comprehension and played a role in their interpretation during the sentence verification task. This relationship is shown in the results, which can be seen in the following chapter. Thus, a participant's average response times to related sentences (true, false, or both), controlling for the participant's overall response speed, can be used as measures of activation of inferencing. To examine if inference activation is different across tasks, a second linear mixed effects model was constructed to predict reaction times to related sentences with correct responses. The fixed effect was task type, and subject was included as a random effect.

3.2.1.2 To what extent does inference generation predict variance in comprehension task outcomes (scores) independent of language ability and individual differences?

To understand whether inference generation differs according to individual and testing factors (question 1b), three LME models were used to predict the dependent variable of reading comprehension score in each of the three task types using the independent variables of inferencing (average response times to related sentences), language proficiency, and individual differences in reasoning, working memory, reading fluency, and motivation as fixed effects, and

random participant effects. The inclusion of inferencing as an independent variable was contingent upon the results of question 1a. It is hypothesized that each task type has a different model of score prediction, with inferencing contributing more predictive power in modeling score of tasks with less response constraint (cloze and summary). Together, these analyses provide insight into the role of inferencing both as a mental product of reading and as a tool in understanding text comprehension.

3.2.2 Research Question 2

3.2.2.1 To what extent does real-time reading behavior, as measured by eye-tracking, differ between reading tasks?

Various statistical methods were also employed to answer the second set of research questions regarding the role of online reading behavior in reading comprehension. Eye-tracking metrics were compared using correlations to identify any measures which were overall pairwise multicollinear, and thus not measuring a distinct enough construct in this dataset. Next, to address this first part of question 2, regarding whether macrotextual reading behaviors differ between task types, eye-tracking measurements are compared for significant differences between the three tasks. Each eye-metric was predicted using linear mixed-effects (LME) regression model with a single fixed effect (Task) and two random effects (individual participant and the six text topics). This was performed using the *lme4* package in R (Bates et al., 2015). R^2 is presented as effect size for each prediction. Only measures with moderate effect sizes were included in the predictive model of tasks. Post-hoc pairwise tests were conducted to understand which of the tasks were significantly different from each other and illustrate the magnitude of each task's effect on eye movement. Previous eye-tracking research suggests verification of statistical results with visual evidence (Kurzhals et al., 2017; Raschke et al., 2014). Thus, in interpreting these

results, visual evidence from scan-paths and heat maps are referenced to provide extra explanation. Finally, a Generalized Logistic Mixed Effects Regression (glmer; Bates et al., 2015) was constructed using eye-tracking metrics as independent variables to predict the dependent variable, task type, controlling again for random individual effects. This type of statistical analysis allows for categorical dependent variables. The ten eye-tracking measurements are transition saccades between text and task, total fixations per word on text and on task, number of fixations per line and paragraph, average duration of fixation on text and on task, average length of saccade, and total rereading time by line and by paragraph. For the logistic regression, the data was split into a training and test set, with 85 participants' three tasks included in a training set to build the model, and the remaining 11 participants datapoints used as a test set to verify the model. Due to the different level of response complexity and required attention to text information, it is hypothesized that higher levels of these measures of text level reading are associated with different tasks.

3.2.2.2 To what extent do online reading behaviors predict variance in reading comprehension scores beyond that predicted by individual differences?

Lastly, to address the second part of question 2, three linear models were constructed to predict the dependent variable of comprehension score in each task type, in these cases using eye-tracking metrics as fixed factors along with predictive individual differences identified as predictive of score in the above-mentioned linear models. Eye-tracking data was split in three sets, one for each reading task. Correlations were calculated between each metric and task score, and further correlations were calculated between each metric and the individual differences. Eye-tracking metrics which were significantly and at least weakly correlated with score, while not

being multicollinear with any other predictor measure, were included in a linear regression model to predict score.

3.3 Summary

In this chapter, I first reported the research questions for the present study. I then detailed the methodology of the study, including information concerning the participants, operationalization of constructs, data collection instruments and procedures, and data preparation. Finally, I provided an overview of the statistical analyses applied to answer each research questions. In the next chapter, I describe the preliminary analyses focused on the validation of the various measures for which data was collected in the above-described procedure.

4 INSTRUMENT RELIABILITY

This chapter presents the various procedures used to measure the reliability and validity of the various scores collected during the data collection procedure. For measures which included discretely scored items, internal reliability was calculated using Cronbach's alpha. These measures include the language proficiency test, logical reasoning test, multiple-choice scores, and cloze scores. For working memory, due to the random nature of the stimulus presentation, and reporting of scores as accuracy percentages, split-half reliability for accuracy on the first and latter halves of the test is calculated instead of Cronbach's alpha. For the motivation survey, a confirmatory factor analysis was conducted to verify that the questions asking about extrinsic and intrinsic motivation factored into two latent variables. For the more subjective summary rating, a full Multifaceted Rasch Analysis was conducted to investigate construct, scale, and intra-rater reliability, and Cohen's Kappa was calculated to measure inter-rater reliability.

4.1 Morpho-syntactic Proficiency

The test used to establish basic L2 proficiency in terms of morpho-syntactic and vocabulary knowledge was scored using a key, and each item was assigned a score of 1, 0.5, or 0. Each participant received a score out of 18. The mean score on the test was 12.573, *sd* = 3.399. Reliability was measured using Cronbach's α, a measure of the internal reliability of the test. It measures the degree to which the individual items on a test correlate with the overall ability of the test-takers. The closer α is to 1, the higher the reliability. The threshold for acceptable reliability is traditionally placed at .7, although shorter tests with fewer participants may have acceptable α below .7. For the proficiency test, Cronbach's α was calculated to be .802.

4.2 Reasoning test reliability

The reasoning test (123test, n.d.) was utilized to measure the inductive reasoning ability of participants. It included 10 dichotomously scored items, presented in order of difficulty, with the first question having easiest intended difficulty. Each participant received a score out of 10. The average score was $M = 7.632$, $sd = 2.409$. Cronbach's α was calculated to evaluate internal reliability of this test as well. For the reasoning test, Cronbach's α was calculated to be .801. Two participants failed to complete the reasoning test and their scores were not reported.

4.3 Working memory test reliability

A 2-back test was employed to measure working memory capacity. It included 35 images, and 15 matches to detect. Correct responses to items as either a match or non-match were recorded, with the final percentage of correct responses used as a score. The average correct response rate was $M = 0.570$ (57%), $sd = 0.237$. One participant failed to complete the working memory test, and their score was not reported.

As the order of stimulus presentation was randomly determined, whether an item was responded to correctly as a match or as a non-match was not aligned for all participants. Thus, internal reliability was calculated using a split-half reliability measurement based on the Spearman-Brown formula, rather than Cronbach's Alpha. In a way similar to Cronbach's alpha, reliability estimates closer to 1 are stronger. Split-half reliability was calculated to be $\rho_{12} = .731$.

4.4 Motivation survey confirmatory factor analysis

The survey used to assess reading motivation utilized 5 items to assess extrinsic motivation and 5 items to assess intrinsic motivation. Each item was responded to in a Likert-scale format from 0 to 4. The questions are presented in Appendix C. The maximum potential

score for each section was 20, indicating high motivation, and the minimum potential score was 0. The sub-surveys for each type of motivation were initially evaluated for internal reliability with Cronbach's α. Reliability of the intrinsic items was satisfactory at $\alpha = .664$, but reliability of the extrinsic items was not sufficient at $\alpha = .250$. Thus, a confirmatory factor analysis (CFA) was carried out to further examine the validity of the survey instrument.

A CFA can be used to determine how well items on a survey relate within the intended constructs. The CFA analysis was completed in R using the Lavaan package (Rosseel, 2012). A mean-adjusted Weighted Least-Squares estimation with robust statistics was employed to examine how well extrinsic and intrinsic motivation could be aggregated from survey data. Tables 1 presents the unstandardized estimates for each item within the two expected latent variables, with standard error, the test statistic and significance of the item's loading into that factor in the pre-test and post-test administrations. At the bottom, this table includes the test statistic of model fit (χ_2), the significance of the fit (p), and the Comparative Fit Index (CFI) and root mean square error of approximation (RMSEA) which compare the fit of the model to the observed data.

From the results in Table 4.4.1, it can be seen that the expected model of constructs was confirmed as a significant factorization of intrinsic items but not of extrinsic items. The model was significant ($\chi_2 = 58.23$, $p = .006$), but RMSEA was a little higher than acceptable at 0.070 and CFI was moderate at 0.886, below the threshold for acceptance of .95. Based on these results, the intrinsic motivation questions can be reliably factored together and used as an aggregate measure of intrinsic motivation, where the extrinsic motivation questions cannot be used as an aggregate measure.

Table 4.4.1 Factor estimates for each survey item in the expected factor.

Latent Variables	Item	Estimate	SE	Wald's *z*	*p*
Extrinsic Motivation					
	E1	1			
	E2	0.759	0.962	0.789	0.430
	E3	0.834	1.064	0.784	0.433
	E4	3.483	3.324	1.048	0.295
	E5	-1.401	1.631	-0.859	0.390
Intrinsic Motivation					
	I1	1			
	I2	0.857	0.200	4.287	< 0.001
	I3	0.575	0.243	2.369	0.018
	I4	0.809	0.222	3.641	< 0.001
	I5	1.884	0.333	5.653	< 0.001

χ_2 = 58.23, p = 0.006, Comparative Fit Index = .886, RMSEA = 0.070

4.5 Comprehension test score reliability

Each set of comprehension task scores was analyzed for reliability in a way that suited the scoring method. Since the MC task was objectively scored by key, Cronbach's α was utilized to measure the internal reliability of each test form. Since the cloze task was objectively scored by key with multiple raters, Cronbach's α was utilized to measure the internal reliability and Cohen's Kappa (weighted) was used to measure inter-rater reliability. Since the summary task was subjectively scored by multiple trained raters using a rubric, a Multi-faceted Rasch Analysis (MFRA) was employed to measure the internal reliability and consistency of the rubric and raters, and Cohen's Kappa (weighted) was used to measure inter-rater reliability. The overall mean scores, score ranges, and reliability metrics are shown in Table 4.5.1, and each reliability analysis is analyzed in depth in the following sections.

Table 4.5.1 Descriptive statistics and reliability for comprehension tasks.

Task	M (SD) Score	Score range	Internal Reliability	Inter-rater reliability
MC	2.854 (1.062)	0 to 4	Average α = .461	N/A
Cloze	9.675 (2.981)	0.75 to 15	Average α = .707	K = .96
Summary (total)	7.699 (2.428)	3 to 12	Rubric construct infit = .99 Rater infit = .95	K = .583

4.5.1 Multiple-choice score descriptive statistics and reliability

MC tasks were scored dichotomously. Each participant received a score out of 5. The overall average score across topics was *M* = 3.287 (*sd* = 1.216) based on the total sample of 102 participants. For each of the six topics, 17 participants took responded to the MC form. Table 4.5.2 presents the internal descriptive statistics for MC scores for each topic, as well as internal reliability for each form.

Text 1 (“Biotechnology”) MC scores had a reliability of α = .378. As this was insufficient reliability, the item which correlated least with the other items was removed, which increased reliability to α = .550. Text 2 (“Compound Microscope”) MC scores had a reliability of α = .035. As this was insufficient reliability, the item which correlated least with the other items was removed, which increased reliability to α = .357. Text 3 (“Water”) MC scores had a reliability of α = .390. As this was insufficient reliability, the item which correlated least with the other items was removed, which increased reliability to α = .402. Text 4 (“Hunger”) MC scores had a reliability of α = .699. This was sufficient reliability, and the removal of any items only reduced reliability. Text 5 (“Choices”) MC scores had a reliability of α = .037. As this was insufficient reliability, the item which correlated least with the other items was removed, which increased reliability to α = .288. Text 6 (“Attitudes”) MC scores had a reliability of α = .134. As this was insufficient reliability, the item which correlated least with the other items was removed,

which increased reliability to α = .472. Overall, although removal of an item increased reliability of each test form, overall reliability was fairly low. This is expected of tests with so few items and this may lower the power of analyses conducted on the MC scores. Adjusted mean scores on MC tests fell within 1 point across topics, and the range was acceptable considering the standard deviations on each topic's scores.

Table 4.5.2 Descriptive statistics and internal reliability for MC tests for each topic.

Text	M	SD	Cronbach's α	Adjusted M	Adjusted SD	Adjusted α
Biotechnology	3.47	1.17	0.378	2.95	1.13	0.550
Microscope	3.78	0.94	0.035	3.11	0.96	0.357
Water	3.18	1.24	0.390	2.71	1.11	0.402
Hunger	3.28	1.53	0.699	2.62*	1.22*	0.699*
Choices	2.39	1.04	0.037	2.50	1.04	0.288
Attitudes	3.61	0.92	0.134	3.22	0.81	0.472

*No item was removed in the MC form for "Hunger". Original scores were scaled to be out of 4.

4.5.2 Cloze score descriptive statistics and reliability

The fifteen-item cloze tests (one for each of six topics) were scored twice using an answer key; once by the researcher and once by a trained rater. Each participant received a score out of 15. Exact agreement across items and participants between the raters and researcher was 91.1%, with a Cohen's Kappa .96. Nevertheless, each disagreement was adjudicated until a single agreed score was assigned to each item for each participant. These adjudicated scores were then used to calculate further descriptive statistics and Cronbach's α for the cloze tests across the 15 items for each topic. The overall average score across topics was $M = 9.67$ ($sd = 2.98$) based on the total sample of 102 participants. For each of the six topics, 17 participants took responded to that topic's cloze form. Table 4.5.3 presents the internal descriptive statistics for cloze scores for each topic, as well as internal reliability for each form. The cloze form for each topic had sufficient internal reliability. Mean scores on cloze tests fell within a 3-point range, which was acceptable considering the standard deviations on each topic's scores.

Table 4.5.3 Descriptive statistics and internal reliability for cloze tests for each text.

Text	M	SD	Cronbach's α
Biotechnology	8.85	3.25	0.763
Compound Microscope	8.47	2.67	0.659
Water	9.47	3.61	0.826
Hunger	10.56	2.55	0.690
Choices	9.36	2.83	0.697
Attitudes	11.43	2.00	0.607

4.5.3 Summary score descriptive statistics and reliability

Each participant completed a summary as one of the comprehension tasks. Summaries were expected to be within 50 to 150 words. Each summary was rated by at least two raters using a rubric developed by the researcher (Appendix F). Raters scored each summary for each of the four constructs on the rubric (Accuracy, Modeling, Task Completion, and Language) on a scale from 0 to 4. If the two first raters for a summary differed by more than 1 point in any of the four constructs, a third rater (and rarely a fourth rater) provided an additional total rating. Average ratings for the closest two scores were used as final scores. Table 4.4 presents descriptive statistics for the summary scores for each topic, across the subscores. In the final column is a total score which adds together the Accuracy, Modeling, and Task Completion scores, while excluding the language score, to give a single summary score based on comprehension out of 12. The cells in Table 4.5.4 show the mean score, with standard deviation in parentheses. Topic 5 was overall scored slightly lower than the other topics, with specifically accuracy being rated lower, and topic 6 was overall scored slightly higher than the other topics, with specifically modeling scores higher than average.

The total comprehension score and each summary component score were compared pairwise with each other. All comprehension components were strongly correlated with each other ($r > .7$) and with total score ($r > .9$). As such, only the total summary score is used as a

dependent variable in subsequent analysis since it captures the overall comprehension construct. Correlations were weaker between comprehension constructs and language, indicating that raters were able to separate, to some degree, the language construct from comprehension. These correlations are shown in Table 4.5.5.

Table 4.5.4 Mean score and standard deviation (sd) for summary scores for each topic.

Text	Accuracy	Modeling	Task Completion	Language	Total Comprehension
Biotechnology	2.74 (.81)	2.53 (.82)	2.62 (.76)	2.38 (.63)	7.88 (2.08)
Compound Microscope	2.94 (.98)	2.29 (.90)	2.53 (.99)	2.65 (1.00)	7.76 (2.74)
Water	2.81 (.84)	2.31 (.97)	2.39 (.99)	2.33 (.71)	7.50 (2.68)
Hunger	2.61 (.85)	2.42 (.73)	2.53 (.92)	2.47 (.74)	7.56 (2.28)
Choices	2.47 (.78)	2.47 (1.01)	2.50 (.95)	2.35 (.79)	7.44 (2.59)
Attitudes	2.76 (.75)	2.65 (.86)	2.50 (.71)	2.59 (.80)	7.91 (2.15)

Note: Total comprehension is calculated as the average sum of Accuracy, Modeling, and Task Completion.

Table 4.5.5 Correlations between summary rubric construct scores.

	Accuracy	Modeling	Task Completion	Language
Modeling	0.736			
Task Completion	0.771	0.841		
Language	0.575	0.649	0.641	
Total	0.900	0.930	0.944	0.673

To investigate the reliability of the rubric constructs, the rating scale, and the raters for scores on the summary forms, a Multi-faceted Rasch Analysis (MFRA) was performed using the program, Facets version 3.83 (Linacre, 2020). This analysis presents a score model for the entire test, which gives information about how well the rubric constructs fit the test model and how well each point on the rating scale differentiated test-takers at different ability levels. It also evaluates the degree to which the raters exhibited self-consistency, or internal reliability. To further investigate the reliability of raters, inter-rater reliability was calculated using Cohen's

Kappa, though the numbers of pairwise ratings between raters was low. Each aspect of reliability on the summary test is detailed below.

Regarding the overall reliability of the summary task to separate examinees at different ability levels, the MFRA had a reported weighted likelihood estimate reliability of .902, indicating high person separation reliability and accuracy of scoring. Infit measures were calculated for each construct on the rubric. For a rubric construct to be reliable, infit measurements should lie within .5 and 1.5 (Linacre, 2002) or ideally within a more narrow range of .8 to 1.3. Infit that is too low indicates that a construct was too narrowly defined and exhibited limited score variance, and infit that is too great indicates a construct was poorly defined and ratings for the construct were erratic, or else did not model well with the other constructs. Fit statistics for each construct on the rubric showed that each construct exhibited sufficient fit and are presented in Table 4.5.6. The higher infit for Language indicates that it was treated by raters in a way inconsistent with the other constructs, meaning it constituted a construct separate from the comprehension constructs. The table also presents the fair average and facility for each construct, indicating that Accuracy was rated highest (the easiest), followed by Modeling and Task Completion, with Language being the lowest rated or most difficult.

Table 4.5.6 Statistics for rubric constructs

Construct	Fair Average	Facility	S.E.	Infit
Accuracy	2.71	-0.59	0.12	1.02
Modeling	2.61	0.01	0.12	0.81
Task Completion	2.48	0.20	0.12	0.85
Language	2.43	0.38	0.13	1.30

Fit statistics were likewise calculated for the rating scale employed by the rubric for each construct. Fit statistics for each scale point on the rubric showed that each point exhibited sufficient fit, and these are presented in Table 4.5.7. A visual presentation of scale functioning

for each construct is further provided in Figure 4.5.1. The charts in Figure 5 show the probability of assignment of a score given an individual's ability level. Each scale point should be represented by a distinct peak, and these peaks should be ordered along the person ability scale in the expected numerical order. Both of these conditions are satisfied by the distributions of score assignment probabilities.

Table 4.5.7 Fit statistics for rubric scale

Scale point	Accuracy Fit	Modeling Fit	Task Completion Fit	Language Fit
4	1.0	1.0	0.9	1.3
3	0.9	0.9	0.8	1.2
2	1.1	0.7	0.7	1.1
1	1.2	0.7	0.8	1.5

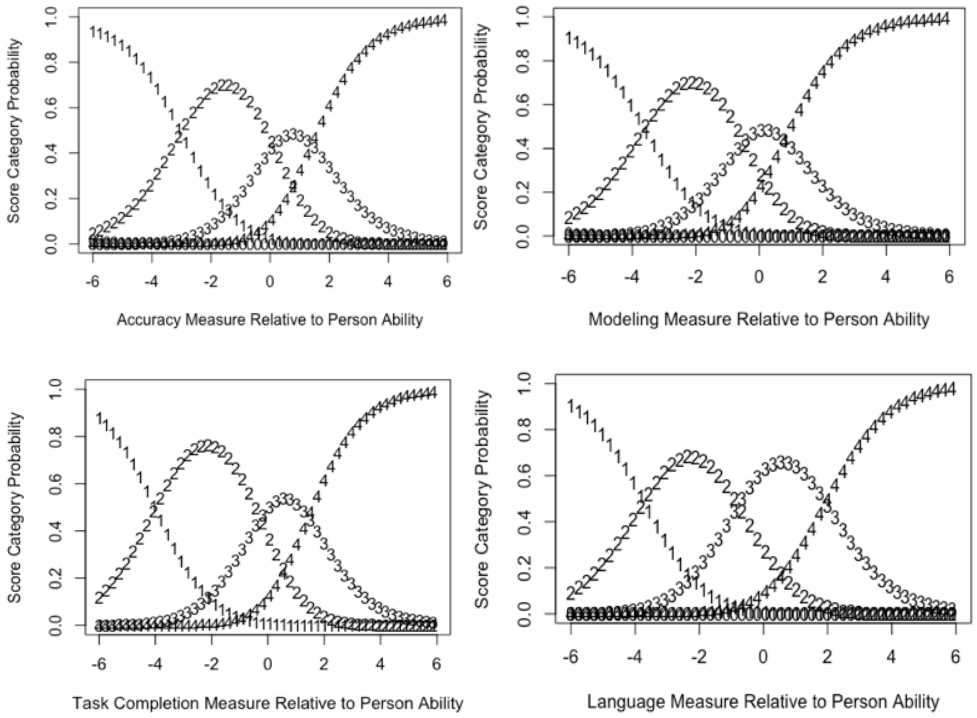


Figure 4.5.1 Summary score probability curves with respect to person ability

Reliability statistics for raters were produced as well. Severity and fit were calculated for each rater to ascertain the degree to which raters differed in overall ratings and exhibited self-consistency. There were seven raters, and their number of summaries rated, severity, fair

average, and fit statistics are presented in Table 4.5.8. The rater separation index was 2.19, with rater separation reliability of .83, indicating that raters exhibited 2.19 distinct levels of severity, and this distinction was significant. This is indicated by two raters being noticeably more lenient than average (Raters G and M) and one rater being noticeably more severe than average (Rater R). Rater G was the most lenient rater (-0.52) and Rater R was the most severe (0.58). Raters exhibited different levels of severity, but no rater's average rating was more than one standard deviation from the mean, indicating raters were neither too severe nor too lenient overall. Tolerable fit has been variously defined as between .5 and 1.5 (Linacre, 2002), .75 to 1.3 (McNamara, Knoch, & Fan, 2019), and .6 to 1.4 (Wright et al., 1994) for rating scales. Taking these bounds into consideration, all raters exhibited satisfactory model fit, indicating self-consistent rating patterns.

Table 4.5.8 Rater statistics

Rater Code	N	Severity	S.E.	Fair average (Total score)	Infit	Point biserial	Exact Agreement
G	24	-0.52	0.18	2.77	0.90	0.79	48.1%
M	24	-0.47	0.17	2.75	0.74	0.80	44.2%
N	44	-0.15	0.13	2.61	1.04	0.80	45.8%
W	18	0.15	0.19	2.48	0.96	0.59	38.2%
I	44	0.19	0.13	2.47	1.10	0.75	47.9%
E	18	0.21	0.20	2.46	0.93	0.76	43.1%
R	34	0.58	0.15	2.32	1.00	0.77	46.2%
Overall inter-rater						0.75	45.5%
Separation Index							2.19
Separation Reliability							0.83

Interrater reliability was further calculated using Cohen's Kappa. This statistic shows the degree to which pairs of raters showed similar trends in rating, and Kappa values closer to 1 are desirable, with values closer to 0, or negative values, indicating poor interrater reliability. Table 4.5.9 presents Cohen's Kappa values for each pair of raters and the number of ratings for each pair. As the number of ratings for a given pair can be quite small, these results must be taken

with caution. Lower sample sizes can influence the accuracy of Kappa values. Two rating pairs, G-I and R-W, showed the lowest interrater reliability, but the raters otherwise exhibited sufficient internal consistency, and low sampling may be the source of lower Kappa values. Considering these results alongside the process for adjudicating disagreement, there is evidence to suppose that summary scoring functioned reliability.

Table 4.5.9 Cohen's Kappa interrater reliability for summary raters.

Rater Pair	N	Cohen's Kappa
E-R	7	.64
E-W	11	.41
G-I	8	.22
G-R	15	.57
I-M	8	.78
I-N	27	.70
M-N	15	.67
R-W	5	.29
Overall	96*	0.583

*Six summaries were set aside as benchmarks for rater training, so the total number of summaries was 102.

4.6 Summary

The current chapter outlined measurements of reliability for the assessment data. This data is utilized in both of the research questions, so ensuring reliability was a critical concern. The results from the reliability statistics indicate that, on the whole, data collection procedures functioned reliably, and where reliability was insufficient, adjustments were made.

5 RESULTS: THE RELATIONSHIP BETWEEN INFERENCING AND SECOND LANGUAGE READING ASSESSMENT

This chapter presents results from the sentence verification task which participants completed after each reading task (i.e., Research Question 1). The first part of this analysis regards how measurements of inferencing based on reaction times differ across sentence types. This inferencing metric is then compared across the different task types and text topics.

The second part of this chapter investigates the relationship between inferencing measurements and reading test scores along with individual differences in reading and language ability. Correlations between inference-generation scores, test scores, and individual differences are calculated and discussed. Correlated variables are then included in a linear model to predict scores on the different tasks. A different model is constructed to predict each different type of reading score outcome in the different tasks. The chapter ends with discussion of the findings, limitations, and directions for future research.

5.1 Research Question 1a: Measuring Inferencing in Reading Assessment

As described in the chapter 3, 102 participants each completed three reading tasks, and each task was followed by a sentence verification task (SVT). As a reminder, the sentence veracity task involved reading a series of 16 sentences. For each sentence, subjects indicated whether the sentence was true or false. The veracity of the sentences was not contingent upon understanding of the texts and were rooted in real world facts or falsities. Eight sentences were related to the text, and eight were irrelevant control sentences. The true/false and related/unrelated categories overlapped, creating a matrix of four sentence conditions: true-related (inferences), false-related (violation of inferences), true-unrelated, and false-unrelated (control sentences). Reaction times to each sentence were calculated as the time it took

participants to make a true or false decision about the sentence. These reaction times are reported in milliseconds per letter (ms/letter) to control for sentences' visual spans. The next subsection reports on descriptive statistics for proportion of correct responses to the sentences and average reaction times for the sentence conditions.

5.1.1 Descriptive statistics for sentence verification task

Data from the Sentence Verification Task (SVT) were first evaluated for correct responses. The purpose of the SVT is to understand how quickly information was accessed in verifying simple sentences, not to understand if the information in the SVT sentences could be drawn successfully from the reading texts, since the content of the verification task sentences was not dependent on the text, but rather general knowledge that could be activated via inferencing during reading. Thus, only response times to correct responses are used for further analysis. Nonetheless, inspection of correct response rates was performed to identify any potential problematic items.

Table 5.1.1 shows descriptive statistics for the rate of correct responses in the SVT for each type of sentence across and between tasks and topics. Across all tasks and topics, true related sentences were accurately responded to 89.6% of the time, and true unrelated sentences were accurately responded to 87.1% of the time. False related sentences had an accurate response rate of 78.1% and false unrelated sentences had an accurate response rate of 78.3%. False sentences in general had a less accurate response rate overall than true sentences, but between sets of related and unrelated sentences, there was not a noticeable difference in correct response rate.

Table 5.1.1 Correct response rates in the Sentence Verification Task

		Related		Unrelated	
		True	False	True	False
		M (SD)	M (SD)	M (SD)	M (SD)
	Total	89.6% (.306)	78.1% (.414)	87.1% (.336)	78.3% (.412)
MC	All Topics	91.1% (.285)	78.% (.415)	85.3% (.355)	77.4% (.418)
	Biotechnology	98.5% (.061)	76.% (.188)	94.1% (.109)	92.6% (.147)
	Microscope	98.5% (.061)	87.7% (.204)	92.6% (.147)	85.8% (.166)
	Water	95.3% (.101)	64.6% (.214)	90.6% (.125)	75.6% (.217)
	Hunger	95.6% (.098)	80.9% (.273)	82.4% (.193)	60.8% (.274)
	Choices	66.2% (.175)	71.6% (.167)	94.1% (.109)	73.5% (.272)
	Attitudes	93.1% (.144)	82.9% (.26)	60.6% (.272)	75.9% (.161)
Cloze	All Topics	87.4% (.332)	80.1% (.4)	86.2% (.345)	78.3% (.413)
	Biotechnology	94.1% (.109)	81.9% (.196)	88.7% (.208)	97.1% (.083)
	Microscope	95.6% (.098)	89.7% (.127)	91.2% (.123)	78.9% (.27)
	Water	85.3% (.178)	76.5% (.225)	92.6% (.147)	84.8% (.182)
	Hunger	84.7% (.152)	69.9% (.274)	81.9% (.224)	56.% (.26)
	Choices	69.3% (.258)	84.4% (.18)	93.8% (.144)	79.7% (.209)
	Attitudes	94.1% (.109)	80.9% (.188)	70.1% (.283)	75.% (.198)
Summary	All Topics	90.1% (.298)	76.1% (.427)	89.6% (.306)	79.3% (.406)
	Biotechnology	92.9% (.144)	66.2% (.259)	96.6% (.098)	85.3% (.218)
	Microscope	97.2% (.081)	86.1% (.154)	97.2% (.081)	78.2% (.181)
	Water	90.3% (.184)	70.4% (.196)	90.3% (.174)	83.3% (.227)
	Hunger	98.3% (.065)	74.4% (.232)	88.3% (.16)	61.7% (.16)
	Choices	67.6% (.23)	77.9% (.239)	98.5% (.061)	78.9% (.184)
	Attitudes	94.1% (.188)	79.4% (.182)	64.2% (.224)	85.3% (.178)

This trend was fairly stable across tasks and topics, although a few sets of sentences had lower correct response rates than the average. The false related sentences for the passage Water, which were also used as the false unrelated sentences for the passage Hunger, had a below-average correct response rate, and upon inspection of the sentences, one sentence ("Water is an element containing multiple smaller molecules.") seemed to account for most of the incorrect response rate skew. Only 20% of responses to this sentence were correct. The true related sentences for the passage Choices, which were also used as the true unrelated sentences for the

Attitudes passage, had a below average correct response rate. Although no one sentence could be identified as problematic, and the sentences were drawn from pieces of information inferable from the passage, the overall tone of the Choices passage is slightly more subjective than the others, and this may have influenced the perceived veracity of the sentences.

Response times to correct responses in the SVT were first checked for outliers. Outliers were found to be any response with less than 18 ms/letter (roughly corresponding to 1 second/sentence) or over 300 ms/letter (roughly 15 seconds/sentence, though this varies). These extreme values were removed from the data set before further analysis.

Due to the slight imbalance in the number of responses per topic used for response time examination, a one-way ANOVA was carried out on response times across the three topics to understand what effect the topics may have on response times to sentences. The ANOVA was found to be significant $F(5, 3997) = 5.772$ ($p < 0.001$), but with a very small effect size (general $eta^2 = .007$). Upon inspection of post-hoc pairwise tests, the only text which had significantly faster reaction times was the Biotechnology text, which had post-reading SVT reaction times significantly faster than the Water text ($p = .008$, $d = .35$), the Hunger text ($p = .004$, $d = .35$), the Choices text ($p = .006$, $d = .35$), and the Attitudes text ($p = .007$, $d = .39$). The effect size was weak in each case[1]. No other pairs of texts had significantly different reaction times overall.

Response times per letter were calculated for each other condition (true/false and related/unrelated). Mean, standard deviations and measures of skew and kurtosis are presented in Table 5.1.2 for all SVT response times, as well as for each sentence condition. As expected in

[1] The heuristic for interpretation of Cohen's *d* considers *d* between .2 and .5 to be a weak effect, between .5 and .8 to be a moderate effect, and above .8 to be a large effect (Cohen, 2013)

response time data, all SVT response times were positively skewed, meaning faster than average response times had lower variance were (i.e. were bunched closely together), and slower than average response times had higher variance with a high ceiling (i.e. were spread farther apart). For all sentence conditions, kurtosis was less than 3 (negative excess kurtosis), indicating that the relative weight of outlying values was not heavy. The following section compares these response times across conditions and tasks using generalized linear mixed effects models, which are more robust against skewed and non-normal data than standard linear models.

Table 5.1.2 SVT response times for different sentence conditions.

Sentence Type	M	SD	Skew	Kurtosis
False-Unrelated	101.701	53.362	1.323	1.567
False-Related	97.493	49.381	1.522	2.453
True-Unrelated	94.006	48.349	1.503	2.679
True-Related	89.990	43.477	1.470	2.724
Total	95.544	48.755	1.463	2.406

5.1.2 Predicting reaction times by sentence types and task conditions

Two generalized linear mixed effects models were constructed to compare response times between sentence conditions. The first model predicted response times with sentence condition, comparing target True-related sentences as a baseline to False-related sentences, True-unrelated sentences, and False-unrelated sentences. The second model included SVT sentence condition and the task of the reading text immediately before the SVT as predictors of reaction time.

The first model treated the True-related sentences as a baseline and other sentence conditions as fixed effects and included participants as a random effect. This yielded a significant model, $F(3, 3696.8) = 12.035$, $p < .001$. This is reported in Table 5.1.3. From the estimates, True-related sentences are responded to significantly faster than to each of the other sentence conditions. Each condition significantly contributed to the model.

Post-hoc inspection of contrasts between each sentence condition revealed that True-related sentences were responded to significantly faster than True-unrelated sentences (t = -2.113, p = 0.035) and significantly faster than False-related sentences (t = -3.441, p = 0.001). However, True-unrelated sentences were not responded to significantly faster than False-related sentences (t = -1.377, p = .169). Thus, in general, the True-related sentence, which were related to ideas inferable in the reading text, were responded to significantly faster than the other conditions, which violated inferences made during the text or were unrelated to the text. The differences in average reaction times between each pair of conditions were analyzed using post-hoc pairwise comparisons, shown in Table 5.1.4.

Table 5.1.3 Linear mixed effects model predicting sentence verification response time using sentence condition

Condition	*B*	B	SE	df	*t*	*p*
Intercept	-0.103	-0.030	0.053	151.46	-1.954	0.052
True-Unrelated	0.084	0.037	0.040	3696.36	2.098	0.036*
False-Related	0.140	0.059	0.041	3696.36	3.388	0.001*
False-Unrelated	0.241	0.102	0.041	3696.36	5.859	< .001*

B = unstandardized coefficients, B = standardized coefficients

Table 5.1.4 Post-hoc analyses of SVT reaction times by condition

Comparison	Mean difference	*t*	*p*
TR – TU	-0.082	-2.098	0.036*
TR – FR	-0.154	-3.388	0.001*
TR - FU	-0.241	-5.859	< .001*
TU - FR	-0.072	-1.339	0.181
TU - FU	-0.158	-3.771	< .001*
FR - FU	-0.086	-2.362	0.018*

Note: TR = True-related, TU = True-unrelated, FR = False-related, FU = False-unrelated

The results of the above model provide evidence that there is a priming effect for the True-related sentences by comparing the reading comprehension tasks on response times to inferable ideas in sentences from the SVT. This provides evidence that response times to related

sentences can be used as a reflection of inference activation during reading. Using reaction times to each correctly responded to related sentence in the SVT and controlling for individual response times, each participant was given an inferencing score. These scores were reversed and scaled so that higher inferencing scores reflected faster average response times to the related sentences controlling for an individual's overall reaction speed.

A second model was constructed to examine the differences in the inference response time metric between tasks. Inference response times were predicted with task type as a fixed effect and subjects and topics as random effects. The baseline task condition was the cloze reading task. This did not yield a significantly predictive model, $F(2, 1818.5) = 1.199$, $p = 0.302$, and the task conditions were not significant predictors within the model. This model is presented in Table 5.1.5. This indicates that the difference in response times to the inferable ideas was stable across sentence verification tasks after each type of reading task.

Table 5.1.5 Linear mixed effects model predicting sentence verification response time using sentence condition and task condition

Condition	*B*	B	SE	df	*t*	*p*
Intercept	95.575		2.563	184.306	37.289	< .001*
MC	-2.539	-0.052	2.382	1816.795	-1.066	0.287
Summary	-3.603	-0.073	2.396	1817.837	-1.504	0.133

B = unstandardized coefficients, B = standardized coefficients

5.2 Research Question 1b: Predicting test score with inferencing and individual differences

This section examines potential connections between reading comprehension scores and sentence verification task response time by examining the effect of the average response time variables for each participant in a linear model to predict score in each of the three tasks. Individual difference measures will also be utilized as predictors of scores. Individual differences

were assessed before the experimental reading tasks, and separate correlations were then calculated for each task condition. Then, for each task, correlations are examined between reading test scores, individual difference measures, and inferencing measurements. The inferencing measurement used in this analysis are the average response times to related sentences. Each participant's average inference response time is controlled for their overall response speed.

After measuring correlations, individual difference measures which were significantly, and at least weakly, related to each task score were used to construct three linear models. The inferencing measure was added as an additional predictor to each model to understand if inferencing, as measured by response times sentences in the SVT related to inferable ideas in the reading text added additional predictive power to the score models. These results are reported below.

5.2.1 Correlations of individual differences

The individual difference measures which were calculated before the reading tasks included a general morpho-syntactic language proficiency score, a reading speed test, a typing speed test, a logical reasoning test, a working memory measurement, and an intrinsic motivation survey. Out of the 102 participants, 2 were excluded for missing data in the individual differences section, one missing a score for reasoning, and one missing scores for reasoning and working memory. For all subsequent analyses in this chapter N = 100. Correlations between the scores for each measure were calculated and are presented in Table 5.2.1. Only one pairwise correlation was significant, with a moderate effect size: reasoning with working memory (r = .358, p < .001). Both measures were retained for further analysis, although variance inflation factors may warrant the removal of variables in subsequent modeling.

Table 5.2.1 Correlations between individual differences

	Morpho-syntactic proficiency	Reading Speed	Typing Speed	Reasoning	Working Memory
Reading Speed	0.100				
Typing Speed	0.105	-0.016			
Reasoning	0.003	0.006	-0.200		
Working Memory	0.139	0.018	-0.171	0.349*	
Intrinsic Motivation	0.218	-0.029	0.078	-0.042	-0.148

N = 102, * $p < .005$

5.2.2 Predicting cloze scores

Correlations related to the cloze task are presented in Table 5.2.2. Scores on the cloze task were strongly correlated with morpho-syntactic proficiency ($r = .630$, $p < .001$) and weakly correlated with reasoning ($r = .212$, $p = .039$) and working memory ($r = .206$, $p = 0.043$). No other measures were correlated with cloze scores, including the post-cloze task inferencing measure from the sentence verification task. The Inferencing Response Time measure was not found to be significantly correlated with any other measures. It was weakly, but negatively, correlated with morpho-syntactic proficiency ($r = -.135$, p = .171). Morpho-syntactic proficiency, working memory, and reasoning were thus included in the baseline linear model to predict cloze scores before adding inferencing response time to the model.

Table 5.2.2 Correlations between measures related to the cloze task

	Inference Response Time	Cloze score
Morpho-syntactic Proficiency	-0.135	0.630*
Reading Speed	-0.083	0.039
Typing Speed	0.047	0.228*
Reasoning	-0.021	0.212*
Working Memory	0.041	0.206*
Intrinsic Motivation	-0.003	0.086
Inference Response Time		0.025

*$p < .05$

The first baseline model to predict Cloze scores included Morpho-syntactic proficiency, typing speed, working memory and reasoning as predictors. Working memory and typing speed showed high variance inflation (i.e. showed nonindependence from reasoning) and were removed from the final model. The final baseline model was found to be significant, $F(2,97) = 39.52$ ($p < .001$). Table 5.2.3 contains a description of the model. The model had a large effect size, explaining about 45.4% of the variance in cloze scores ($r_2 = .454$). Both reasoning and Morpho-syntactic proficiency were found to be significant predictors of score in the model, with Morpho-syntactic proficiency being the stronger predictor based on standardized coefficients (also see correlation in the previous section).

The inference response time measure (average response time to related sentences) was included as a predictor in a second model. Since inferencing ability has been found to be different between higher and lower proficiency readers in previous studies (Feller et al., 2020; Lake, 2014; Shimizu, 2009), interaction between proficiency and inferencing was also included as a predictor. When inference response time was added to the model, the model remained significant, $F(4,95) = 20.38$ ($p < .001$), with a similarly large effect size ($r_2 = .467$). Neither the inference measure nor the interaction with proficiency were significantly predictive in the model, and proficiency and reasoning remained significant. This model is shown in Table 5.2.4.

Table 5.2.3 Linear regression model predicting Cloze scores

Condition	*B*	B	SE	*t*	*p*	r_2	Δr_2
Intercept	-0.119		1.231	-0.097	0.923		
Morpho-syntax	0.573	0.634	0.070	8.169	< .001*	0.397	
Reasoning	0.339	0.226	0.123	2.758	0.007*	0.454	0.057

B = unstandardized coefficients, B = standardized coefficients, * $p < .05$

Table 5.2.4 Linear regression model predicting Cloze scores including inferencing

Condition	*B*	B	SE	*t*	*p*	r_2	Δr_2
Intercept	-0.367		1.250	-0.293	0.770		
Morpho-syntax x Inference RT	-0.003	-0.003	0.005	-0.559	0.578	0.004	
Inference RT	0.062	0.255	0.068	0.924	0.358	0.021	0.017
Morpho-syntax	0.603	0.668	0.072	8.387	< .001*	0.410	0.393
Reasoning	0.355	0.237	0.113	3.139	0.002*	0.467	0.057

B = unstandardized coefficients, B = standardized coefficients, * $p < .05$

5.2.3 Predicting MC scores

Correlations related to the multiple-choice task are presented in Table 5.2.5. Scores on the multiple-choice reading task were weakly correlated with reasoning ($r = .221$, $p = .025$). No other measures were significantly correlated with MC score, though there was a weak yet insignificant correlation with Morpho-syntactic proficiency ($r = .177$, $p = .095$) and working memory ($r = 0.189$, $p = .063$). The inference response time ratio calculated from the SVT after the MC task was not significantly correlated with MC score or any other measures. Morpho-syntactic proficiency, reasoning, and working memory were thus included in the first linear model to predict MC score before adding inferencing response time to the model.

Table 5.2.5 Correlations between measures related to the MC reading task

	Inferencing Response Time	MC score
Morpho-syntactic proficiency	0.033	0.177
Reading Speed	-0.124	0.032
Typing Speed	-0.029	0.212*
Reasoning	0.044	0.221*
Working Memory	0.047	0.189
Intrinsic Motivation	-0.048	0.079
Inferencing Response Time		0.107

*$p < .05$

The baseline model to predict MC scores included Morpho-syntactic proficiency, typing speed, working memory, and reasoning as predictors. Working memory and typing speed were found to have high variance inflation (due again to collinearity with reasoning, the stronger predictor). Morpho-syntactic proficiency was not found to be a significant predictor and was removed from the baseline model. The final baseline model with only reasoning as a predictor was found to be significant, $F(1,99) = 5.204$ ($p = 0.025$). A description of this model can be found in Table 5.2.6. The model had a small effect size, explaining about 5.1% of the variance in MC scores ($r_2 = .051$). When inference response time ratio was added to the model, the model was no longer significant, $F(4,95) = 2.955$ ($p = 0.057$), with a small effect size ($r_2 = .059$). The inference measure did not significantly contribute to the model. This model is shown in Table 5.2.7.

Table 5.2.6 Linear regression model predicting MC scores

Condition	*B*	B	SE	*t*	*p*	r_2
Intercept	2.098		0.362	5.797	< .001*	
Reasoning	0.117	0.227	0.051	2.281	0.025*	0.051

B = unstandardized coefficients, B = standardized coefficients, * $p < .05$

Table 5.2.7 Linear regression model predicting MC scores including inferencing

Condition	*B*	B	SE	*t*	*p*	r_2	Δr_2
Intercept	2.109		0.363	5.814	< .001*		
Reasoning	0.115	0.223	0.051	2.243	0.027*	0.051	
Inferencing RT	0.088	0.085	0.104	0.848	0.398	0.059	0.008

B = unstandardized coefficients, B = standardized coefficients, * $p < .05$

5.2.4 Predicting summary scores

Summaries were rated for Accuracy, Modeling, Task Completion, and Language (see the rubric in Appendix F). The total summary score was calculated as the sum of the component

scores, excluding the language score. Correlations between total summary score, the summary task inferencing measure and individual differences are presented in Table 5.2.8. Total summary score was weakly to moderately correlated with Morpho-syntactic proficiency (r = .297, p = .003), moderately correlated with Intrinsic Motivation (r = .342, p = .001), and weakly correlated with inference response time ratios (r = .214, p = 0.029). The correlation between score and inference response time ratio was negative, indicating participants' having relatively quicker response times to the inferable sentences in the SVT is related to higher summary scores. No other correlations were significant between summary scores or inferencing response time ratio with other individual difference measures. Inferencing response time showed a weak but insignificant correlation with Morpho-syntactic proficiency, indicating a possible interaction between the two in the metrics' relation to summary score. Summary total score was correlated with Morpho-syntactic proficiency and intrinsic motivation, so these two measures were included in the first linear model to predict summary score before adding inferencing response time to the model.

Table 5.2.8 Correlations between measures related to the summary task

	Inferencing Response Time	Summary score
Morpho-syntactic proficiency	0.125	0.297*
Reading Speed	-0.056	0.036
Typing Speed	0.077	-0.008
Reasoning	-0.064	0.086
Working Memory	-0.101	-0.048
Intrinsic Motivation	0.099	0.345*
Inferencing Response Time		0.214*

*p < .05

The baseline model to predict total Summary scores included Morpho-syntactic proficiency and intrinsic reading motivation as predictors and was found to be significant, F(2,97) = 10.801 (p < .001). Table 5.2.9 contains a description of the model. The model had a

moderate effect size, explaining about 18.5% of the variance in summary scores (r^2 = .185). Both Morpho-syntactic proficiency and intrinsic motivation were found to be significant predictors of score in the model with comparable predictive power.

Table 5.2.9 Linear regression model predicting Summary scores

Condition	*B*	B	SE	*t*	*p*	r^2	Δr^2
Intercept	5.296		0.798	6.634	< .001*		
Intrinsic Motivation	0.189	0.293	0.061	3.083	0.003*	0.122	
Morpho-syntax	0.623	0.258	0.229	2.718	0.008*	0.185	.063

B = unstandardized coefficients, B = standardized coefficients, * $p < .05$

Inference response time was added to the model, and as before, interaction with proficiency was also included in the model. When inference response time ratio was added to the model, the model remained significant, F(4,95) = 6.499 ($p < .001$), with a larger effect size (r^2 = .219). The inference measure was a significant predictor in the model, along with proficiency and motivation, though no interaction effect was observed. Unlike in the score models for the other tasks, the inclusion of the inferencing response time measure was significant and significantly increased the r^2 of the model. This model is shown in Table 5.2.10.

Table 5.2.10 Linear regression model predicting Summary scores including inferencing

Condition	*B*	B	SE	*t*	*p*	r^2	
Intercept	5.432		0.793	6.849	< .001*		
Inference RT x morpho-syntax	-0.071	-0.110	0.236	-0.301	0.764	0.021	
Inference RT	0.447	0.186	0.225	1.990	0.049*	0.057	0.036
Intrinsic Motivation	0.179	0.276	0.061	2.928	0.004*	0.163	0.106
Morphosyntax	0.585	0.242	0.228	2.563	0.012*	0.219	0.056

B = unstandardized coefficients, B = standardized coefficients, * $p < .05$

5.3 Discussion

5.3.1 Summary and connection to previous research

In this chapter, the results from the sentence verification task (SVT) were examined to answer the first research question of whether inferencing during realistic second language reading testing could be captured in a post-hoc task, and whether results from this measure contributed to predicting scores on the reading tests. The SVT was presented to participants three times with three different sets of sentences, and each time it was presented, it followed a different reading test task: reading with multiple-choice questions, reading in a cloze task, and reading while summarizing. In these tasks, participants were shown 16 sentences and responded with a true or false response by button press. Four of the sentences were true and related to the reading text presented just before the SVT without directly repeating information from the passage, while being inferable from the text. Four sentences were true, but unrelated directly to the previous passage. Four sentences were related to the passage without directly repeating passage information, but contained a false element, thus violating inferences generated during passage reading. Four sentences were both false and unrelated to the passage. As these true and false statements could be evaluated using general or background knowledge, accuracy was not the target measurement, but rather the response time to those sentences. Only sentences which participants responded to correctly were included in analyses of response times.

5.3.1.1 Research Question 1a

The hypothesis of the first part of this research question is that response times to the inferable sentences would be faster on average than to unrelated sentences. This is because the information needed to respond correctly to the related sentences, whether previously known or not, would be activated by reading the passage, whereas reading the passage would not activate

information in the unrelated sentences and impact response time. Overall, average response times to the different sentence types followed the expected pattern, with sentences related to inferable information responded to faster than unrelated sentences. The sentences were compared using a linear mixed effects model, and the differences in response times were significant. This difference was stable across the task conditions of the reading passages. This implies that the activation of inferable information while passage reading occurs regardless of the reading goals set by the comprehension task. Thus, the question remained as to whether the activation of inferable information related significantly to success on the different reading tasks.

5.3.1.2 Research Question 1b

To investigate the impact of real-time inference generation on comprehension scores, the average response time to related sentences was calculated for each participant, controlling for the participant's overall response speed, to operationalize the activation of inferable ideas during reading for each task condition. Since participants completed three SVTs, they received three inference response time scores; one for each reading task. Inferencing response time was used in addition to other individual differences to predict scores on the different reading tasks. The individual difference measures were Morpho-syntactic proficiency, reading speed, reasoning, working memory, and intrinsic motivation.

Before constructing predictive score models, correlations between measures were calculated. Correlations were calculated between each task score and the predictor measures. Cloze scores, while significantly but weakly correlated with reasoning and working memory, were very strongly correlated with Morpho-syntactic proficiency. This is likely somewhat inflated by the fact that Morpho-syntactic proficiency was measured using a gap-fill test variation, which is similar to cloze tasks in format. Beyond the superficial similarity of the tasks,

there is reason to believe that the cloze task does require more lower-order lexical and morpho-syntactic knowledge to complete, so the strong correlation with Morpho-syntactic proficiency is reflective of the understanding of cloze tests in previous research (J. C. Alderson, 2000; Raatz & Klein-Braley, 1981). In linear modeling, cloze scores were predicted by Morpho-syntactic proficiency to a large extent, with reasoning also contributing significantly to the linear model. Inference response time ratio did not significantly correlate and was marginally significant in the model to predict cloze scores. Although there is evidence that cloze texts as reading tests are primarily affected by L2 proficiency, we will return to understanding more about what predicts cloze success in the next chapter.

MC reading task scores were significantly but weakly correlated with Morpho-syntactic proficiency, reasoning, and working memory, and the correlation with reasoning was slightly stronger than the others. The linear models created to predict MC score showed that reasoning alone was predictive of score, albeit with a small predictive power. Neither Morpho-syntactic proficiency nor working memory contributed significantly to the models. The inclusion of inferencing to the model rendered the model insignificant, with no real change in predictive power. These results make sense since responding to MC tasks requires evaluating and eliminating answer choices with respect to reading topics. The finding that a nonverbal reasoning measure most closely relates to MC task success aligns with previous research which indicates that MC reading tasks run the risk of relying on surface-level and macro-strategies related to deciphering and analyzing answer choices rather than modeling text (Khalifa & Weir, 2009; Rupp et al., 2006). Although a connection was found with reasoning, the models predicting MC score were the weakest of all models, indicating a large amount of unaccounted variance in

scores. This could imply reading for MC questions is a unique construct, or else dependent on yet unexplored variables. This is investigated further in the next chapter.

For summary tasks, summary score was correlated with Morpho-syntactic proficiency, intrinsic motivation, and inference response time. In linear models predicting scores, Morpho-syntactic proficiency, motivation, and inference response time each contributed predictive power in modeling summary score. The relationship of summary scores and intrinsic motivation is well-attested, as previous research found links between intrinsic motivation and general reading comprehension (Guthrie et al., 2007; Wigfield & Guthrie, 1997) and specifically with the construction of more meaningful and complex summaries (Fransson, 1984). The relationship between summary scores with L2 proficiency is understandable given the correlation of proficiency with the other test tasks, but it is unclear if proficiency is directly impacting reading comprehension, or if productive language ability played a role in the summary scores even though language ability was rated separately and not included in the above scores. Taylor (2013) warns that open-ended summary production tasks put extra linguistic demands on test takers (p. 72), which may account for the predictive power of Morpho-syntactic proficiency on the overall summary score.

Unlike with MC and cloze scores, inference response time ratio was found to be correlated significantly, but weakly, with summary scores, indicating that responding relatively faster to inferable sentences was related to more successful summary writing. In addition, inference response time ratio was significantly predictive of modeling scores, contributing a change in r_2 of .036. This was the only model in which inferencing contributed to score prediction. Otherwise, the inference response time measure was overall a weak predictor of scores and weak correlate with other abilities.

This specific finding is reasonable considering the explicit modeling component of summary writing and rating (see the rubric in Appendix F). Van Dijk & Kintsch's (1977) and Brown and Day's (1983) models for summarization each include the selection, exclusion, and superordination of information in a source text, all of which are demanded by the summarization task. Superordination, or subsuming multiple ideas into more general ideas, requires inferencing ability to read across a text and fill in gaps necessary to condense information. This highlights the role of inferencing in text comprehension: inferencing is critical for the activation of schema (Anderson & Pearson, 1984) and making causal links for building a mental representation of a text (van den Broek et al., 2015). The modeling construct on the summary rubric was designed to capture test takers' mental model construction, and to the extent that summary score correlated with a key subskill of mental modeling, the rubric appeared to be successful in capturing this process.

The complete linear model of summary scores also aligns with previous research on schemata and inferences in L2 reading. Nassaji (2002) concluded from a survey of research on L2 reading and schema theory that L2 readers devote more resources to efficient decoding of texts than activating inferencing, even when they have demonstrated inferencing ability. This is not to say that readers lack inferencing ability, but instead that it secondary in importance to Morpho-syntactic proficiency in predicting comprehension. Likewise, inferencing response times in this study are not a discrete measure of inferencing ability as used in research on inferencing in L2 reading (Feller et al., 2020), but rather an attempt to capture inference generation as it occurred during realistic reading assessment task completion. In this regard, the current study's linear model to predict summary modeling confirms this understanding, showing inferencing as a significant predictor of summary score, but not as strong a predictor as proficiency.

Although the measure of inferencing was predictive of summary outcomes, the correlation between score and inferencing was stronger than the linear unidirectional relationship. There is still the chance that, rather than inferencing ability contributing directly to modeling scores, the test takers' reading purpose, to write a well-modelled summary, pushed for greater activation of inferencing in test-takers who wrote higher-rated summaries. This particular strength of summary writing was seen in Caccamise et al (2007), who found that the task of summary writing induced more active reading and situation model building than other reading tasks.

5.3.2 Conclusions and implications

In conclusion, regarding the first part of research question 1, whether examinees respond significantly faster to sentences inferable from a text than to unrelated sentences after reading a text, the results from this study show that related sentences are responded to significantly faster than other types of sentences in a sentence verification task following passage reading. This difference in response speed is not dependent on the type of reading task completed during passage reading. Regarding the second part of research question 1, the extent to which inference generation predicts variance in comprehension task outcomes (scores) independent of Morpho-syntactic proficiency and individual differences, the results from this study show that A) inference generation only impacts reading outcomes when the measured reading score is explicitly designed around an aspect of reading where inferencing is critical (i.e. mental modeling) and B) the impact of inferencing on scores is secondary to that of Morpho-syntactic proficiency, reasoning and intrinsic motivation when it is predictive of scores.

For test design, the findings of this study provide evidence that higher-order reading skills can be captured in L2 reading tests if desirable, but the aspect of mental modeling must be

explicitly built into design and scoring of the test. Each reading test task examined in this chapter had significant predictive models with unique sets of predictors. Thus, it is important for reading tests to utilize a variety of tasks to account for the many subskills which contribute to academic second language reading. Specifically, if the goal of a reading test is to capture a higher-order cognitive reading skill such as inferencing, the summary task with clear guidelines for evaluation based on mental modeling is most likely to capture this skill.

5.3.3 Limitations and future directions

There are several areas of limitation in this study, and subsequently many avenues for further research. The study highlights the difficulty in examining inferencing in expository texts, which are more information dense, put more responsibility on the reader to interpret information, not necessarily linear, and require more specific background information for comprehension when compared to narratives, the type of text usually employed to understand inference generation (Lorch, 2015). In L1 reading literature, reading expository texts has been found to be more likely to trigger literal comprehension processes and discourage unnecessary inferencing past those necessary for local coherence (Noordman et al., 1992). Noordman et al. (1992) further assert that inferencing during expository text may be dependent upon goal setting, a conclusion for which the current study provides some support. For more precise understanding of inferencing in academic L2 reading, further research is needed in general on inference generation while reading expository texts.

The current study also makes no practical distinction between the various types of inferences which could be made during reading, such as bridging inferences, causal inferences, or elaborative inferences, instead treating inferencing as a general ability to insert default or logical information into comprehension gaps. Although there is support for examining

inferencing as a general skill (Kendeou, 2015), further research may examine specific types of inferences which can be drawn from expository texts to understand if specific types of inferencing contribute to comprehension.

Although data for many measures related to reading and language ability were collected for this study, one type of data which could not be collected was direct L1 literacy data. Due to the diverse pool of participants, an L1 literacy measure would be unfeasible. By examining correlates of literacy, such as reasoning, working memory, and motivation, it was hoped that skills which may contribute to successful L1 literacy could be captured, but this is not guaranteed. As previous models of L2 reading generalize the components of L2 reading to be either L2 proficiency- or L1 literacy-based (Koda, 1988), it is difficult to situate the results of these findings. L2 proficiency was certainly found to be related to reading comprehension, more-so than other individual differences, but a comparison between L2 proficiency and a general literacy ability could not be compared here. Future studies may include L1 reading comprehension tests to create a fuller picture of the skills which contribute to L2 reading comprehension.

From a methodological standpoint, there are several limitations. Although over 100 participants were recruited for this study, this is still a relatively small sample size considering the types of analyses conducted, especially after accounting for outliers and missing data. The linear models to predict reading task scores may suffer from low power. Post-hoc power analyses based on real effect sizes found the average power of the linear models in this study to be around 61% on average, less than anticipated in the a priori analysis. Thus, the chance for false negatives are fairly high, and future studies taking a similar approach to understanding inferences in L2 reading will require larger sample sizes.

The sample size limitation influenced other aspects of the statistical analyses. For instance, interaction effects and topic effects were not included in score models due to the inability for the sample size to sustain so many predictor variables. Thus, nuanced investigations of how predictor variables may interact and create thresholds for activation of other variables were not possible. Further research can include interaction effects, or else remove a continuous variable, such as reasoning or proficiency, from models, and create separate models for discrete groups or bands within these variables.

Another quantitative limitation was the number of sentences supported by the SVT. As the expository texts used in this study were fairly short, information dense, and targeted toward those with little background knowledge on a given subject, there were few opportunities to isolate inferable ideas from the texts, and thus few data points to rely on for each sentence condition for each text. Future studies can utilize this method with longer source texts and a larger pool of related sentences.

Additionally, regarding the use of SVT in this context, previous uses of SVTs have typically been used to understand differences between types of stimuli and experimental conditions. In this regard, the current study contains findings of this type, with sentences related to the priming reading passage responded to faster than unrelated sentences. However, there is less use of SVT to understand within person differences, and the task may be less suited for this purpose. It is thus unsurprising that the relationship between the post-hoc SVT and reading comprehension performance was weak overall.

Inference generation is a critical aspect of text modeling and reading comprehension, but it is not the only skill which can provide evidence of global text processing and complex text modeling. The next chapter takes a different approach to understanding reading comprehension

tasks and score outcomes by investigating real-time reading behavior data from eye-tracking methods. This approach may provide insight that a post-hoc measure could not capture, such as evidence of strategic reading, attention, specific fluency, and global integration of textual information.

6 RESULTS: THE RELATIONSHIP BETWEEN EYE-MOVEMENT BEHAVIOR AND SECOND LANGUAGE READING ASSESSMENT

This chapter presents results related to the second research question outlined in chapter 3, *To what extent does real-time reading behavior, as measured by eye-tracking, differ between reading tasks, and to what extent do online reading behaviors predict variance in reading comprehension scores beyond that predicted by individual differences?* This question is related to eye-movement behavior data gathered from participants during each reading task. The first part of this chapter reviews the eye-movement data collection procedure, explains the eye-tracking metrics used, and provides descriptive statistics for the eye-tracking measures.

The second section presents findings related to the first part of the research question comparing various eye-tracking metrics between the three reading tasks. Although eye-tracking provides a plethora of data for interpreting the reading process, the eye-tracking metrics utilized for this study are those that are most comparable between the reading tasks. These include total number of fixations per word while reading the text, the average fixation duration while reading the text and while interacting with the task area, the average number of fixations per dwell

The third part of this chapter presents findings related to the second part of the research question and investigates the relationship between eye-tracking metrics, reading test scores and other individual differences in reading and language ability. Correlations between eye-tracking metrics, test scores, and individual differences are calculated and discussed. Correlated variables are then included in a linear model to predict scores on the different tasks. A different model is constructed to predict each different type of reading score outcome in the different tasks. The chapter ends with a discussion of the findings, limitations, and directions for future research.

6.1 Overview of methods

6.1.1 Description of eye-movement measures

The purpose of this chapter is to present results from analyses of real-time reading behaviors between reading comprehension tasks and understand to what degree these behaviors influence comprehension score outcomes. This involved examination of aggregated eye movement data gathered from participants while they completed the reading comprehension tasks. Visual data from participant scan paths were examined to ensure recorded data aligned with the image presented onscreen and did not include too many erratic and unexpected fixations[2]. Unaligned data occurred in the case of two of the 102 participants, whose data needed to be discarded due to poor alignment between eye-tracking recording and screen captures. Aggregated heatmaps for each text-topic combination are presented in Appendix G. These maps provide a general display of the relative intensity of attention across the text in each condition.

The data used for the analyses in this chapter were calculated using participants' fixation location and duration data, and further analyzed based on whether fixations and saccades took place within predefined AOIs (see chapter 3). The eye-tracking metrics described in chapter 3 which were calculated in this study are summarized in Table 6.1.1. The following section gives an overview of the methods employed to analyze the eye-tracking data.

6.1.2 Methods and analyses

Several statistical methods are employed in this chapter to investigate eye-movement behavior during second language reading assessment. First, descriptive statistics for each eye-tracking measurement in each task and topic were gathered.

[2] Although each participant's scan-path was manually examined by the researcher to ensure quality, this data is too cumbersome to present here, and aggregated measures are shown in Appendix G,

Table 6.1.1 Description and operationalization of eye-tracking measures

Measure	Purpose for measurement	Target area	Operationalization notes
Fixations on text per word	Global, careful reading	Entire text area	Average of all fixations made on the reading text in a given trial.
Mean length of saccade	Global reading	Entire trial area	Average absolute distance between sequential fixation coordinates throughout a trial.
Mean duration of rereading dwells in lines/ paragraphs	Global reading	Line/paragraph areas of interest	Mean length of dwells after the first pass through areas of interest. Some scattered fixations may occur early in the trials, before reading has truly begun, and these sporadic fixations are not counted as first passes or toward rereading.
Mean fixations per line dwell	Linear, local reading	Line areas of interest	Average count of fixations per dwell across dwells in line AOIs. Controlled for number of words in AOI.
Mean fixations per paragraph dwell	Local, careful reading	Paragraph areas of interest	Average count of fixations per dwell across dwells in paragraph AOIs. Controlled for number of words in AOI.
Mean duration of fixations on text	Careful reading	Entire text area	Average time (ms) of fixations in any text area of interest. Controlled for size of AOI.
Mean duration of fixations on task	Careful reading, Task integration	Task areas of interest	Average time (ms) of fixations in any task area of interest. Controlled for size of AOI.
Fixations on task per word	Task integration	Task areas of interest	Average of all fixations made on the task areas in a given trial. Size of the areas in the respective tasks is controlled for.
Number of gaze transitions between text and task	Task integration, global reading	Text and task areas of interest	Raw count of saccades which moved from a text area of interest to a task area of interest.

As text topic is not a primary concern in this study, eye-tracking metric means were compared using one-way ANOVA to observe any topic effects in each condition. Additionally, skew and kurtosis data was calculated for each metric to ensure the normality of each measure in each task condition. The above analyses are not reported in detail and were merely performed to ensure the assumptions were met for subsequent analyses. These results are shown in Appendix H.

Eye-tracking metrics were then compared using correlations to identify any measures which were overall pairwise multicollinear, and thus not measuring a distinct enough construct in this dataset. Of any measures which were multicollinear, the measure with a larger effect size difference between tasks was retained for further analysis.

In section 6.2 of this chapter, a series of analyses were conducted between reading tasks, within participants, to compare eye-tracking metrics. The primary goal is to establish whether certain eye-movement behaviors are predictive of the types of reading motivated by the purpose of the reading task. This began with a comparison of eye-tracking metric differences between tasks based on linear mixed effects and pairwise comparisons. Each eye-tracking metric was predicted using a single fixed effect, task, and two random effects (individuals and the six topics). Post-hoc pairwise tests were conducted to understand which of the tasks were significantly different from each other and illustrate the magnitude of each task's effect on eye movement. Only measures with moderate effect sizes were included in the predictive model of tasks. A generalized logistic mixed effects regression model was constructed to predict task type using eye-tracking metrics. The model included one dependent variable, reading task, and included as fixed effects any eye-tracking metrics found to be significantly different between the tasks with at least a small effect size, excluding those found to be variance inflating factors. Individual participant and text topic were included as random effects. The effect size of the model is pseudo r_2, and the predictive power of the model is compared to a baseline model's chance of identifying the reading task by chance (.333 repeating).

In section 6.3 of this chapter, to address the second part of question 2, three linear models were constructed to predict the dependent variable of comprehension score in each task type, in these cases using eye-tracking metrics as fixed factors along with predictive individual

differences identified as predictive of score in the above-mentioned linear models. Eye-tracking data was split in three sets, one for each reading task. Correlations were calculated between each metric and task score, and further correlations were calculated between each metric and the individual differences found to be predictive of score in the previous chapter: Morpho-syntactic proficiency and reasoning for the MC and cloze scores, and Morpho-syntactic proficiency and intrinsic motivation for summary scores. From these correlations, eye-tracking metrics which were significantly and at least weakly correlated with score, while not being multicollinear with any other predictor measure, were included in a linear regression model to predict score. These models again utilized individual difference measures previously found to be predictive to compare the influence of eye-tracking metrics on score relative to individual differences.

6.2 Correlations and selection of eye-tracking metrics

Eye-tracking metrics were compared using pairwise correlations to check for multicollinearity between metrics. Metrics which were found to be correlated at r = +/- .7 or more extreme were considered to be multicollinear, i.e. so closely related that they essentially measure the same underlying construct. Table 6.1.5 reports the correlations between eye-tracking metrics for the entire dataset. Several metrics were significantly correlated, but most metrics did not exhibit multicollinearity. However, both metrics for rereading were multicollinear with each other and with fixations per word on text and fixations per word on task. This entails that the rereading metrics calculated in this data were essentially equivalent (measured at line level or paragraph level) and did not add additional information beyond total number of fixations made. This indicates either that a great amount of rereading was necessary across tasks in this reading setting, or that a more sophisticated approach to calculating rereading may be necessary. This idea is returned to in the discussion section of this chapter. The fixation per word metrics for text

and task were less correlated with the other metrics overall, so the two metrics for rereading were considered the less potentially explanatory variables and thus not included in further analyses.

Table 6.2.1 Correlations between eye-tracking metrics

	A.	B.	C.	D.	E.	F.	G.	H.	I.
A. Mean Length of Saccade									
B. Transitions	**0.442**								
C. Fixations per word (Text)	-0.049	**0.542**							
D. Mean Fixation Duration (Text)	-0.103	0.063	**0.300**						
E. Mean Fixation per Dwell (by line)	**-0.224**	-0.065	0.053	**0.329**					
F. Mean Fixation per Dwell (by paragraph)	**-0.346**	-0.161	**0.222**	**0.343**	**0.536**				
G. Mean Fixation Duration (Task)	**0.316**	**0.463**	**0.287**	**0.235**	-0.005	-0.122			
H. Fixations per word (Task)	0.017	**0.530**	**0.699**	**0.288**	-0.052	0.182	**0.333**		
I. Average Rereading Duration (per line)	-0.081	**0.466**	*0.951*	**0.376**	0.060	**0.269**	**0.290**	*0.716*	
J. Average Rereading Duration (per paragraph)	-0.061	**0.467**	*0.918*	**0.344**	-0.083	0.183	**0.271**	*0.708*	*0.930*

Note: Bold correlations are significant at $p < .001$. Correlations in italics signify multicollinearity.

In total, eight eye-tracking metrics were found to be normally distributed and non-multicollinear with other eye-tracking variables. These include: Fixation per word on text AOIs, Fixation per word on task AOIs, Mean fixation duration on text AOIs, Mean fixation duration on task AOIs, Mean fixation per dwell on text line, Mean fixation per dwell on text paragraph, Mean length of saccade, and number of text-to-task transitions. The following section reports results for the first part of research question 2 using these metrics.

6.3 Research Question 2a: Comparing eye movement behavior between reading tasks

This section reports results for comparisons of eye-tracking metrics between the three reading tasks: MC questions task, cloze task, and summary task. The eight eye-tracking metrics retained from the analyses reported in the previous section were each compared between tasks

using linear mixed-effects regression model predicting each eye-tracking metric using a single fixed effect, task type, and two random effects of topic and individual. Post-hoc analyses were then performed to identify the direction of the effect between the three tasks. The eye-metrics which were significantly different between tasks with a moderate effect size were then used as predictors in a logistic regression to predict task.

6.3.1 Task effects on eye movement measures

A linear mixed-effect regression model was constructed for each eye-tracking metric with task as a fixed effect and individuals and topics as random effects. The full model for each metric is reported in Appendix I. A summarized description of the effect of each model is presented in Table 6.3.1. Each metric was significantly predicted by task, with significant models for each metric at $p < .001$. Effect sizes for each comparison were also examined[3]. Using this heuristic, each comparison resulted in a significant model with at least a weak marginal effect size for task on each eye-tracking metric, and in some cases, larger effect size.

To understand the specific pairwise differences, post-hoc paired t-tests were used for each task on each metric. These results are found in Table 6.3.2, and a visual plotting of means for each metric can be found in box plots in Appendix J. Effect sizes for differences were calculated using Cohen's d. Effect sizes greater than .5 are interpreted as moderate, and those greater than .8 as large. There was at least one significant pairwise difference between reading tasks for each eye-tracking metric calculated. Most effect sizes for significant differences were large, indicating a strong effect for task upon the type of reading behavior elicited. Results of these pairwise comparisons speak to the nature of the reading performed during the three

[3] In regression modeling, A weak r^2 is considered anything above 0.02, a moderate r^2 is considered anything above .09, and a large r^2 is considered anything greater than .25 (Cohen, 2013)

comprehension tasks. In the discussion of this chapter, these results are compared to visual data from eye-tracking scan-paths to paint a clearer picture of the comprehension process in each task.

Table 6.3.1 Summary of linear models predicting eye-tracking metrics with tasks

Eye-tracking metric	Model significance	Marginal r_2	Conditional r_2
Mean length of saccade	F(2,189) = 63.449, p < .001	r_2 = 0.217	r_2 = 0.516
Transitions	F(2,189) = 54.515, p < .001	r_2 = 0.225	r_2 = 0.334
Fixations per word (text)	F(2,189) = 76.345, p < .001	r_2 = 0.254	r_2 = 0.531
Mean text fixation duration	F(2,179) = 32.661, p < .001	r_2 = 0.065	r_2 = 0.723
Mean fixation per line dwell	F(2,181) = 11.665, p < .001	r_2 = 0.034	r_2 = 0.598
Mean fixation per paragraph dwell	F(2,185) = 36.037, p < .001	r_2 = 0.129	r_2 = 0.494
Mean task fixation duration	F(2,190) = 406.04, p < .001	r_2 = 0.682	r_2 = 0.761
Fixations per word (task)	F(2,188) = 66.342, p < .001	r_2 = 0.302	r_2 = 0.352

Note: Full model descriptions are presented in Appendix I

Although there are distinct differences between the tasks regarding each eye-tracking metric, it remains to be seen whether these differences in reading behavior are distinct enough to be unique to and predictive of the reading comprehension tasks. The following section explores this, presenting results from a logistic regression to predict reading task based on eye-movement behaviors.

6.3.2 Logistic regression to predict reading task

A generalized logistic mixed effects regression (GLMER) modeling method was used to predict a categorical variable using several continuous variables and random effects. In this case, the logistic regression model involves predicting reading task based on variation in the eight eye-tracking metrics, controlling for within participant variance and topic variance. To ensure that

the model is not overfit to the sample and was predictive, the model was validated using leave-one-out cross-validation. In this way, a model is built using every instance (i.e., a reading comprehension task performance) except for one, and the model is then used to predict the task of the left-out instance using the relative importance of eye-tracking metrics in the GLMER model. Predictions for each instance are recorded and compared to the actual task for each instance in a confusion matrix to assess the overall accuracy of model predictions (see below).

Table 6.3.2 Post-hoc comparisons for eye-tracking metrics between tasks.

Measure	Comparison	Difference of means	p	d
Fixations per word (text)	Cloze – MC	1.583	< .001	1.209
	Cloze – Summary	0.636	< .001	0.485
	Summary – MC	0.947	< .001	0.723
Fixations per word (task)	Cloze – MC	4.243	< .001	1.344
	Cloze – Summary	2.320	< .001	0.735
	Summary – MC	1.923	< .001	0.609
Mean Length of Saccade	Cloze – MC	13.004	0.005	0.409
	Cloze – Summary	-22.492	< .001	0.708
	Summary – MC	35.496	< .001	1.117
Transitions	Cloze – MC	32.923	0.002	0.440
	Cloze – Summary	-57.452	< .001	0.768
	Summary – MC	90.375	< .001	1.208
Mean fixation duration (text)	Cloze – MC	0.017	< .001	0.584
	Cloze – Summary	0.015	0.001	0.501
	Summary – MC	0.002	0.824	0.083
Mean fixation duration (task)	Cloze – MC	0.107	< .001	0.985
	Cloze – Summary	-0.110	< .001	1.013
	Summary – MC	0.217	< .001	1.998
Mean fixation per dwell (by line)	Cloze – MC	-0.034	0.005	0.454
	Cloze – Summary	-0.013	0.426	0.178
	Summary – MC	-0.021	0.127	0.277
Mean fixation per dwell (by paragraph)	Cloze – MC	0.024	0.039	0.334
	Cloze – Summary	0.060	< .001	0.853
	Summary – MC	-0.037	< .001	0.519

The initial model included all eight of the variables compared above, but average task fixation duration and mean length of saccade were found to have high Variance Inflation Factors, meaning their covariance with other variables prevented unique predictive power. These were removed from the model. The remaining six eye-movement metrics were used to construct the task predicting model, resulting in a significant model on the training set. Three models were initially constructed, each with a different ordering of the tasks (i.e. which tasks were predicted by smaller, moderate, or larger model values). The most accurate of the three models is presented in Table 6.3.3. This lists the included factors, their coefficients, log odds, standard error, chi-square, and significance in the model. The coefficients (B) show the direction of prediction for factors in the model. Negative predictions were associated with the cloze task, and positive predictions were associated with the summary task, with MC predictions in between. The log odds column shows how many times more likely an instance was likely to be classified given a standard deviation change in the metric. For example, a single standard deviation change in fixations per word in the text would make a prediction of cloze 2.478 times more likely. The effect size was calculated using McFadden's pseudo-r^2, which was high at .586.

Table 6.3.3 Generalized logistic mixed effects model to predict reading tasks using eye-tracking metrics

Predictor	B	*Log odds*	SE	χ^2	*p*
(Intercept)	3.977		1.192	4.289	0.038
Transitions	0.073	5.450	0.013	31.960	< .001
Fixation per word (Text)	-1.890	-2.478	0.370	26.018	< .001
Fixation per word (Task)	-0.813	-2.569	0.147	30.430	< .001
Mean Fixation Duration (Text)	-18.039	-0.536	8.601	4.399	0.036
Mean Fixation per dwell (line)	22.379	1.663	4.043	7.857	0.005
Mean Fixation per dwell (paragraph)	-11.332	-0.802	4.766	22.047	< .001
Pseudo R^2	0.586				

Cross validation indicated that the prediction error of the model ($\delta = .084$) was lower than the baseline chance model ($\delta = .223$). The accuracy of the leave-one-out task classification modeled by combinations of eye-tracking metrics is presented in table 6.3.4 in the form of a confusion matrix. The model's overall accuracy in the test set was 58.19%, against a baseline chance of correction prediction of 33.33%. This is significantly more predictive than baseline, $\chi_2 = 182.153$, $p < .001$.

Table 6.3.4 Confusion matrix for logistic regression predictions of task type.

	Predicted task			
Actual task	Cloze	MC	Summary	Accuracy
Cloze	66	22	7	69.47%
MC	3	14	79	14.58%
Summary	7	2	87	90.63%
Overall % Correct				58.19%

Note: Overall % correct by chance = 33.33%

6.3.3 Summary

In this section, results from comparisons of eye-tracking metrics between three types of reading texts and from modeling of those texts using eye-tracking measures were presented. The results together show that each task elicits a different set of reading patterns as evidence by eye movement behaviors.

Results from the logistic regression indicate that the association between eye-tracking metrics and reading tasks goes beyond associated mean differences, and the relationship between eye-movement behavior and reading task is strong enough to predict task using eye-tracking metrics. The prediction is not perfect however, and the misclassification of MC tasks as summary tasks indicates that there is still overlap between the reading behavior activated by readers' goal setting.

As the different reading behaviors have been shown to differ significantly between reading tasks, the investigation shifts to understanding how these reading patterns impact reading performance. The following section reports results of comparisons between the above-described eye-tracking metrics and reading comprehension scores in each task.

6.4 Research Question 2b: Using eye movement behavior to predict reading scores

Section 6.3 covers the results related to connections between eye-movement behavior and reading comprehension performance. Considering the differences in eye-movement behavior by task as reported in the previous section, comprehension scores for each task were predicted with separate models. For each task, this begins with examining correlations between eye-tracking metrics and reading scores. The purpose will be to decide which eye-tracking metrics to include in linear regressions to predict comprehension scores. Thus, correlations between eye-tracking metrics and individual differences shown in chapter 5 to significantly predict score will also be calculated. The metrics which showed a significant correlation with score and at least a weak effect size, while not being multicollinear with any other variable, were included in linear regression models to predict score. Only significant predictors were left in the final models. These results are described below.

6.4.1 Predicting cloze scores using eye movement metrics

Pearson's *r* was calculated for each eye-tracking metric within the cloze task data. Table 6.4.1 shows results from correlations of eye-tracking metrics on the cloze task. It additionally shows correlations with cloze score, as well as Morpho-syntactic proficiency and reasoning, which were found to be significant in the previous chapter. Three metrics were significantly and at least weakly correlated with score: transitions (r = -.207), mean fixation duration on text (r = -.306), and number of fixations per word on task (r = -.212). Each of these was negatively

correlated with score. Fixations per word on task and transitions were almost perfectly correlated (r = .967), and each was multicollinear with fixations per word on text. Since fixations per word on task has a slightly stronger correlation with score than transitions, only mean fixation duration on text and fixations per word on task will be used for modelling of cloze score. Neither of the eye-tracking metrics were strongly related to Morpho-syntactic proficiency or reasoning, the individual difference metrics previously found to predict cloze score, indicating they are independent of other variables predictive of score.

Table 6.4.1 Correlations between eye-tracking metrics in the cloze task

Measure	A.	B.	C.	D.	E.	F.	G.	H.
A. Mean Length of Saccade								
B. Transitions	**-0.297**							
C. Fixations per word (Text)	***-0.377***	*0.854*						
D. Mean Fixation Duration (Text)	-0.158	***0.414***	**0.361**					
E. Mean Fixation per Dwell (by line)	**-0.232**	**0.225**	**0.236**	**0.210**				
F. Mean Fixation per Dwell (by paragraph)	***-0.458***	**0.312**	**0.299**	**0.279**	**0.506**			
G. Mean Fixation Duration (Task)	-0.088	***0.544***	***0.387***	***0.594***	**0.306**	0.166		
H. Fixations per word (cloze gap)	**-0.270**	*0.967*	*0.831*	***0.407***	**0.215**	**0.304**	***0.509***	
Morpho-syntactic proficiency	**0.234**	-0.121	-0.136	-0.141	0.119	-0.090	-0.002	-0.116
Reasoning	-0.047	**-0.220**	-0.136	-0.092	-0.142	-0.085	-0.081	**-0.221**
Cloze Score	0.198	**-0.207**	-0.144	**-0.306**	0.195	-0.132	-0.109	**-0.212**

Note: After applying Bonferroni Correction, correlations in bold and italics were significant at $p < .001$. Correlations in bold were of at least a weak effect size and at least significant at standard $p < .05$. Correlations with italics only are multicollinear.

To check whether interactions effects might affect the modeling, scores on the cloze task were plotted as a factor of each predictor index, and the resulting best fit lines were used as a visual guide for identifying interactions. The participants were split into groups for above median or below median in proficiency, and likewise for reasoning, to make the plots reader friendly. This grouping is not used in further analysis.

Figure 6.4.1 shows visually the plotting of cloze scores along the y-axis, with mean text fixation duration along the x-axis, and line groupings for relative Morpho-syntactic proficiency level and reasoning level. The different slopes of the mean text fixation duration fit lines between proficiency levels and reasoning levels indicates there may be an interaction effect between these three variables. Figure 6.4.2 shows the plotting of cloze scores along the y-axis, with cloze gap fixations per word along the x-axis, and line groupings for relative Morpho-syntactic proficiency level and reasoning level based on a high-low median split. The different slopes of the cloze gap fixations per word fit lines between reasoning groups indicates a potential interaction between these two variables. As such, these interactions were included in the linear modeling.

Since interaction effects are being considered and the predictor variable are on different orders of magnitude, variables were standardized before being entered into the model. Thus, only standardized coefficients are presented. The linear regression model developed for cloze score used as predictors Morpho-syntactic proficiency, reasoning, mean duration of fixation on text, and fixations per word on cloze gaps. Proficiency, reasoning and mean text fixation duration were found to be significant predictors and were kept. Fixations per word on cloze gaps was found to have a high variance inflation factor, and it was thus removed from the final model. Additionally, no pairwise interaction effects were significant in the original model and were thus removed from the final model.

A three-way interaction between Morpho-syntactic proficiency, reasoning, and mean fixation duration was significant and included alongside main effects. The final model was found to be significant, $F(4,94) = 27.64$ ($p < .001$). Table 6.4.2 contains a description of the model. The model had a large effect size, explaining about 55.9% of the variance in cloze scores ($r^2 = .559$), which is more predictive than the model with only Morpho-syntactic proficiency and reasoning ($r^2 = .470$). The significant three-way interaction between Morpho-syntactic proficiency, reasoning, and mean text fixation duration is complex, but the positive coefficient of this interaction indicates that when levels of any two of the predictors rise, the third is more likely to become positively predictive of score.

As is seen in Figure 6.4.1 above, when proficiency and reasoning are both above average, the relationship between score and mean text fixation duration is no longer negative, but positive. This relationship is captured in the model. Main effects for reasoning and Morpho-syntactic proficiency remained significant predictors of score in the model, with Morpho-syntactic proficiency being the strongest predictor based on standardized coefficients. Mean fixation duration had a significant main effect on cloze score with a negative coefficient and predictive power similar to reasoning based on change in r-squared and standardized coefficients.

6.4.2 Predicting MC scores using eye movement metrics

Table 6.4.3 shows results from correlations of eye-tracking metrics on the MC task. It additionally shows correlations with MC score, as well as reasoning, which was found to be a significant predictor of MC score in the previous chapter[4]. Three metrics were significantly and at least weakly correlated with score: transitions ($r = -.293$), mean fixation duration on the

[4] L2Morpho-syntactic proficiency was not found to be predictive of MC score in previous models (see chapter 5) and was not included in modeling here.

question area (r = -.379), and number of fixations per word on questions (r = -.318). Each of these was negatively correlated with score, and none were multicollinear with other eye-tracking metrics or reasoning in the MC task data. Each of the three variables will be used for modelling of MC score.

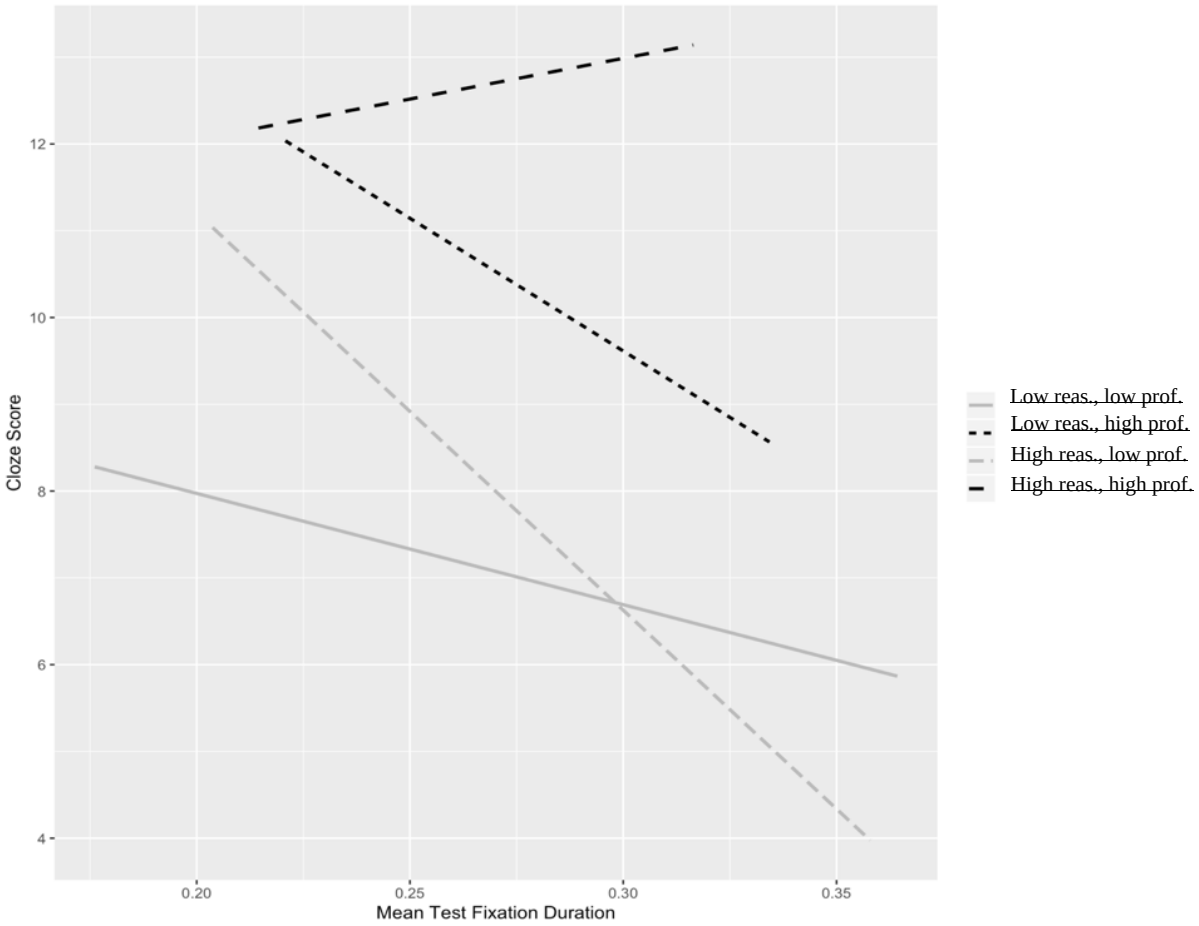


Figure 6.4.1 Cloze score plotted against mean text fixation duration, with groupings for above-median and below-median proficiency and above-median and below-median reasoning.
Note: reas. = Reasoning, prof. = Morpho-syntactic proficiency

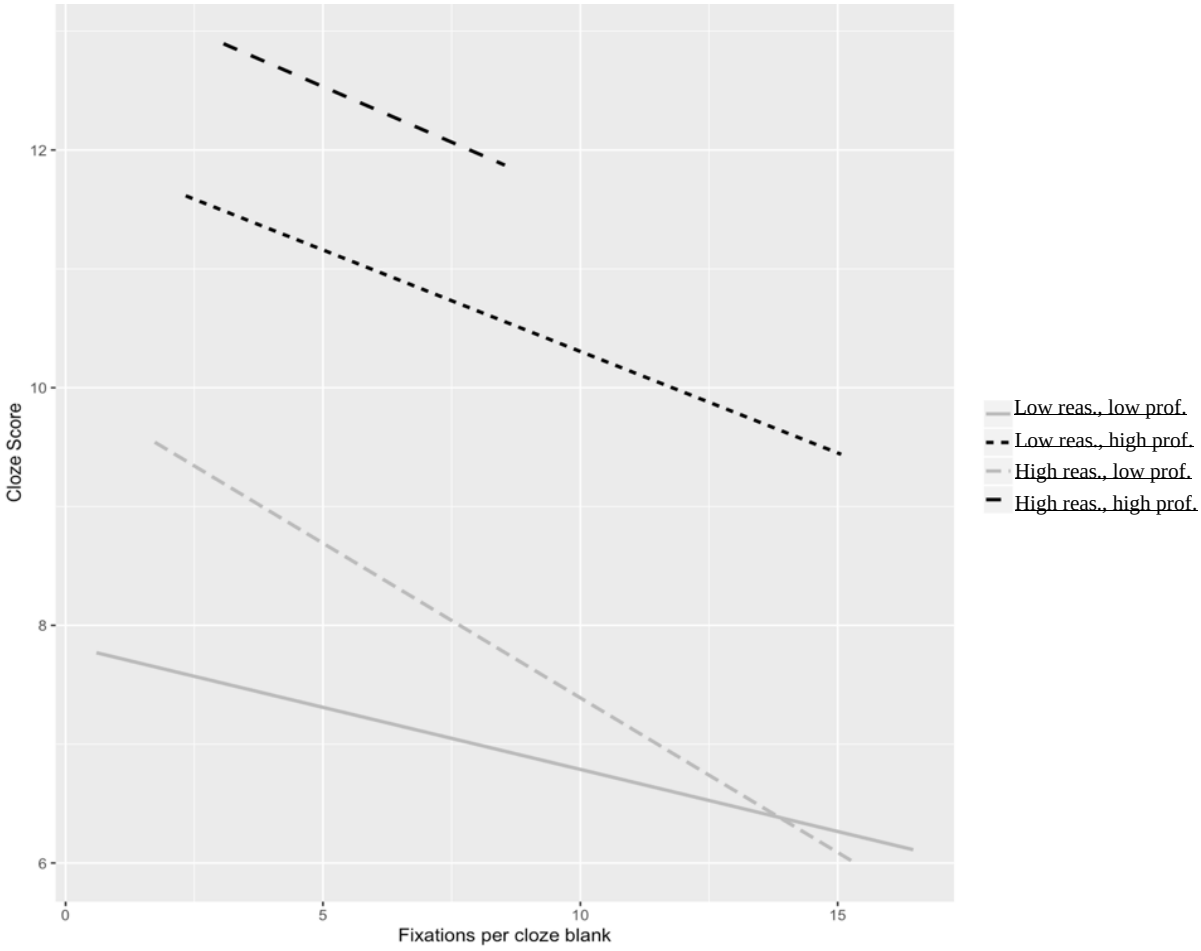


Figure 6.4.2 Cloze score plotted against cloze gap fixations per word, with groupings for above-median and below-median proficiency and above-median and below-median reasoning.
Note: reas. = Reasoning, prof. = Morpho-syntactic proficiency

Table 6.4.2. Linear regression model to predict cloze task scores

Predictor	B	SE	*t*	*p*	r_2	Δr_2
Intercept	-0.021	0.072	-0.293	0.770		
Morphosyntax x Reasoning x Mean text fixation duration	0.151	0.070	2.150	0.034*	0.021	
Morphosyntax	0.663	0.074	8.959	< .001*	0.442	0.421
Reasoning	0.278	0.078	3.551	0.001*	0.511	0.069
Mean text fixation duration	-0.222	0.072	-3.099	0.003*	0.559	0.048

B = standardized coefficients, * significant at $p < .05$

Table 6.4.3 Correlations between eye-tracking metrics in the MC task

Measure	A.	B.	C.	D.	E.	F.	G.	H.
A. Mean Length of Saccade								
B. Transitions	0.173							
C. Fixations per word (Text)	-0.169	**0.537**						
D. Mean Fixation Duration (Text)	**-0.289**	0.031	0.159					
E. Mean Fixation per Dwell (by line)	**-0.423**	-0.129	**0.213**	**0.486**				
F. Mean Fixation per Dwell (by paragraph)	**-0.424**	**-0.295**	**0.202**	**0.322**	**0.653**			
G. Mean Fixation Duration (Task)	**-0.300**	0.072	0.115	**0.615**	0.209	0.084		
H. Fixations per word (Task)	-0.131	**0.586**	**0.643**	**0.214**	0.141	0.098	**0.377**	
Reasoning	0.001	-0.111	-0.015	0.010	-0.170	-0.109	0.010	-0.160
MC Score	0.037	**-0.293**	-0.163	-0.174	0.006	0.137	**-0.379**	**-0.318**

Note: After applying Bonferroni Correction, correlations in bold and italics were significant at $p < .001$. Correlations in bold were of at least a weak effect size and at least significant at standard $p < .05$. Correlations with italics only are multicollinear.

To check whether interactions effects might affect the modeling, scores on the MC task were plotted as a factor of each predictor index, and the resulting best fit lines are used as a visual guide for identifying interactions. The participants were split into groups for above median or below median in reasoning to make the plots reader friendly. This grouping is not used in further analysis.

Figure 6.4.3 shows the plotting of MC scores along the y-axis, with mean task fixation duration along the x-axis, and line groupings for relative reasoning level. The similar slopes of the mean text fixation duration fit lines between reasoning levels indicates there is likely no interaction effect between the variables. Figure 6.4.4 shows the plotting of MC scores along the y-axis, with number of transitions along the x-axis, and line groupings for relative reasoning level. The similar slopes of the transitions fit lines between reasoning groups indicates there is

likely no interaction. Figure 6.4.5 shows the plotting of MC scores along the y-axis with fixations per word on questions along the x-axis, and line groups for relative reasoning level. The different slopes of the fit lines indicate there is a potential interaction between reasoning and fixations per word on questions. Thus, this interaction was included in the linear model.

Since interaction effects are being considered and the predictor variable are on different orders of magnitude, variables were standardized before being entered into the model. Thus, only standardized coefficients are presented. The linear regression model was developed for MC score using as predictors reasoning, mean text duration of fixation on text, and fixations per word on questions. Task fixations per word was found to have a high variance inflation factor and was removed from the model. Of the remaining predictors, mean task fixation duration and transitions were found to be significant predictors, whereas reasoning and interactions with reasoning were not. These effects were thus removed from the model. The final model was found to be significant, $F(2,96) = 12.583$ ($p < .001$). Table 6.4.4 contains a description of the model. The model had a moderate effect size, explaining about 21.3% of the variance in MC scores ($r^2 = .213$). Mean fixation duration on questions was the most significant predictor based on standardized coefficients, with shorter fixations on questions contributing to higher scores. Transitions were also a significant predictor, with fewer transitions predictive of higher score.

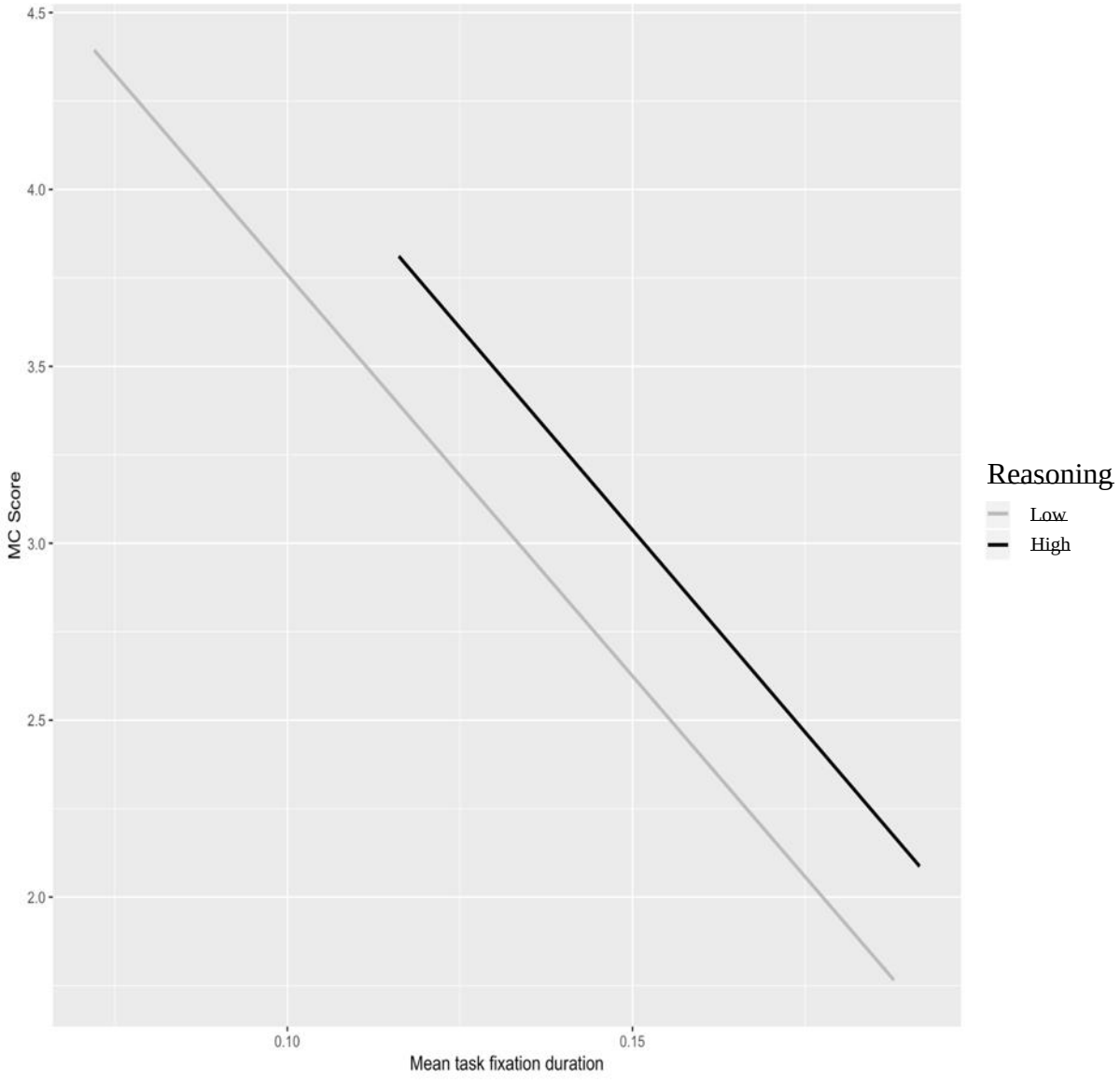


Figure 6.4.3 MC score plotted against mean task area fixation duration, with groupings for above-median and below-median reasoning.

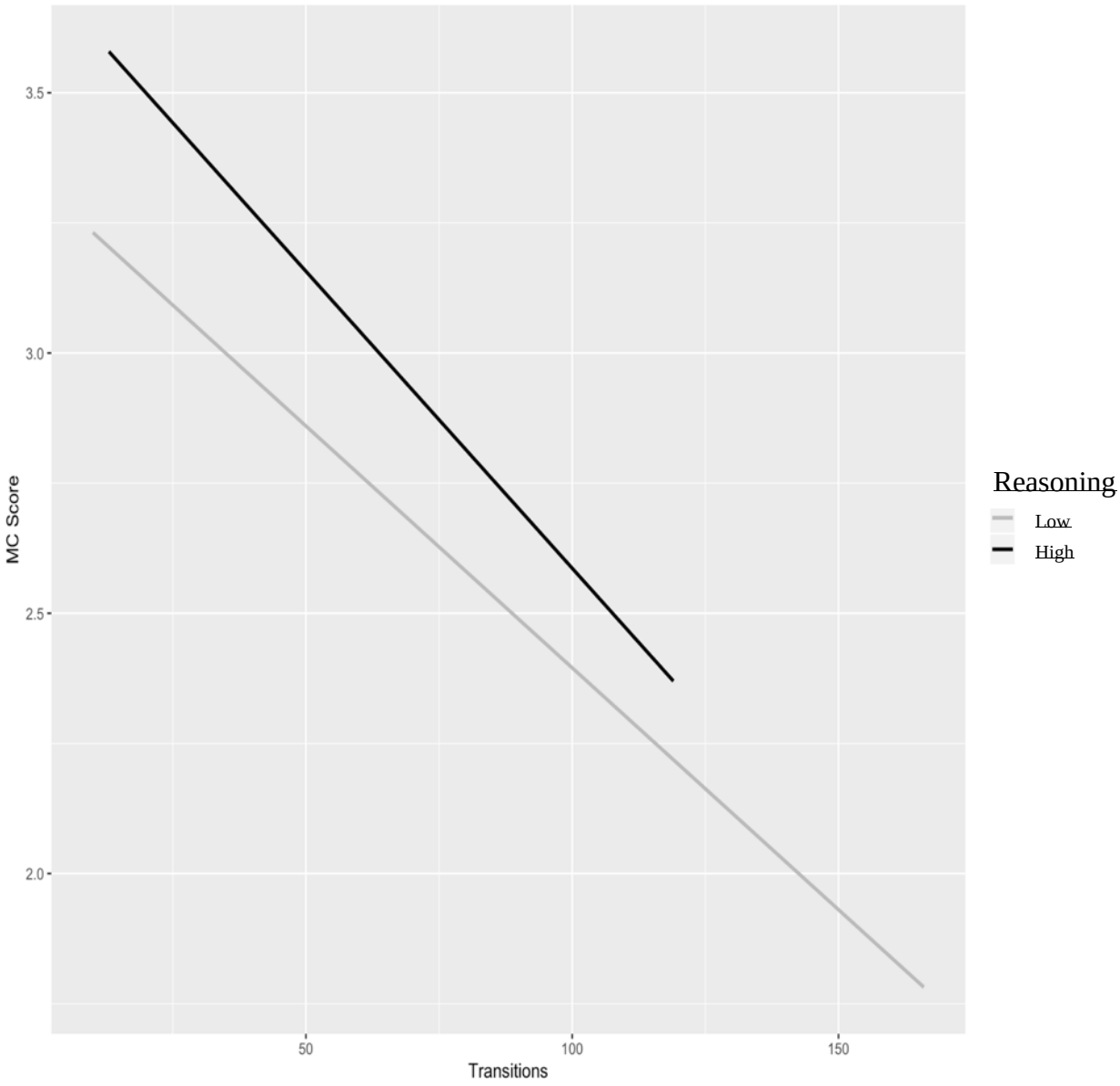


Figure 6.4.4 MC score plotted against number of transitions, with groupings for above-median and below-median reasoning.

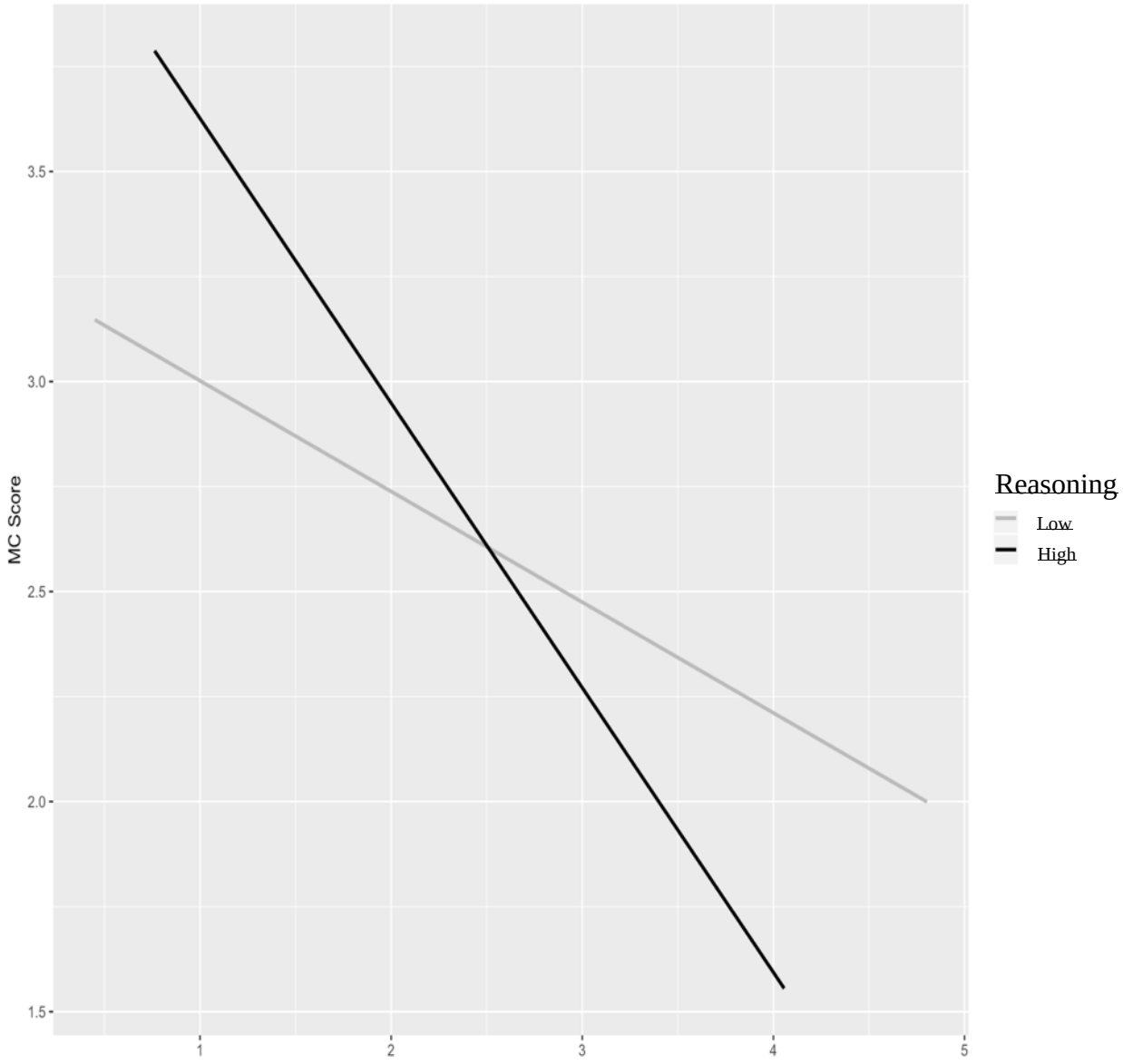


Figure 6.4.5 MC score plotted against fixations per word on questions, with groupings for above-median and below-median reasoning.

Table 6.4.4 Linear regression model to predict MC task scores

Predictor	B	SE	*t*	*p*	r_2	Δr_2
Intercept	< .001	0.092	0.000	1.00		
Mean Fixation Duration (Task)	-0.347	0.092	-3.766	< .001*	0.135	
Transitions	-0.280	0.092	-3.036	0.003*	0.213	0.078

B = standardized coefficients, *significant at p < .05

6.4.3 Predicting summary scores using eye movement metrics

Table 6.4.5 shows results from correlations of eye-tracking metrics on the summary task. It additionally shows correlations with total summary score (the sum of the summary subscores Accuracy, Modeling, and Task Completion), as well as Morpho-syntactic proficiency and intrinsic motivation, which were found to be significant in the previous chapter. Three metrics were significantly and at least weakly correlated with score: transitions (r = .302), fixations per word on the reading passage (r = .364), and mean fixation duration on text (r = -.214). Contrary to the correlations in the cloze and MC data, number of transitions was positively correlated with score in the summary data. Fixations per word on the text was also correlated with summary score. However, as in the cloze data, mean duration of fixations on the text was negatively correlated with summary score. None of these eye-tracking metrics were strongly related to Morpho-syntactic proficiency or intrinsic motivation, the individual difference metrics previously found to predict summary score.

To check whether interactions effects might be present in the modeling, scores on the summary task were plotted as a factor of each predictor index, and the resulting best fit lines are used as a visual guide for identifying interactions. The participants were split into groups for above median or below median in reasoning to make the plots reader friendly. This grouping is not used in further analysis. Figure 6.4.6 shows the plotting of summary scores along the y-axis, with text fixations per word along the x-axis, and line groupings for relative motivation and Morpho-syntactic proficiency level. Figure 6.4.7 likewise shows the plotting of summary scores along the y-axis, with mean text fixation duration along the x-axis, and line groupings for motivation and proficiency level. Figure 6.4.8 shows summary scores along the y-axis, with number of transitions along the x-axis and line groupings for motivation and proficiency level

based on a high-low median split. For all three eye-tracking metrics, the intersecting slops of the fit lines between groups indicate there is possibly an interaction effect between individual difference and eye-tracking variables. Thus, all possible interactions were included in the linear model.

Table 6.4.5 Correlations between eye-tracking metrics in the summary task

Measure	A.	B.	C.	D.	E.	F.	G.	H.
A. Mean Length of Saccade								
B. Transitions	***0.420***							
C. Fixations per word (Text)	-0.150	***0.477***						
D. Mean Fixation Duration (Text)	-0.063	-0.101	0.057					
E. Mean Fixation per Dwell (by line)	-0.171	-0.107	0.115	***0.491***				
F. Mean Fixation per Dwell (by paragraph)	**-0.228**	**-0.266**	0.091	***0.400***	*0.794*			
G. Mean Fixation Duration (Task)	-0.121	-0.040	-0.165	**0.243**	0.108	0.080		
H. Fixations per word (Task)	0.001	***0.484***	**0.267**	-0.065	-0.098	-0.154	0.098	
Morpho-syntactic proficiency	0.135	0.060	0.056	-0.162	0.078	**0.217**	-0.100	-0.121
Intrinsic Motivation	-0.044	-0.045	0.115	0.002	-0.037	0.057	-0.045	-0.167
Summary Score	0.066	**0.302**	***0.364***	**-0.214**	0.023	0.049	-0.132	-0.165

Note: After applying Bonferroni Correction, correlations in bold and italics were significant at $p < .001$. Correlations in bold were of at least a weak effect size and at least significant at standard $p < .05$. Correlations with italics only are multicollinear.

Since interaction effects are being considered and the predictor variable are on different orders of magnitude, variables were standardized before being entered into the model. Thus, only standardized coefficients are presented. The linear regression model was developed for summary score using as predictors intrinsic motivation, Morpho-syntactic proficiency, text fixations per word, mean text fixation duration, and number of transitions. Transitions and its interactions with

individual differences were not found to be significant to the model and were removed. The final model was found to be significant, $F(6, 92) = 9.641$ ($p < .001$). Table 6.4.6 contains a description of the model. The model had a large effect size, explaining about 39.7% of the variance in summary scores ($r_2 = .397$), which is more predictive than the model with only Morpho-syntactic proficiency and motivation ($r_2 = .164$).

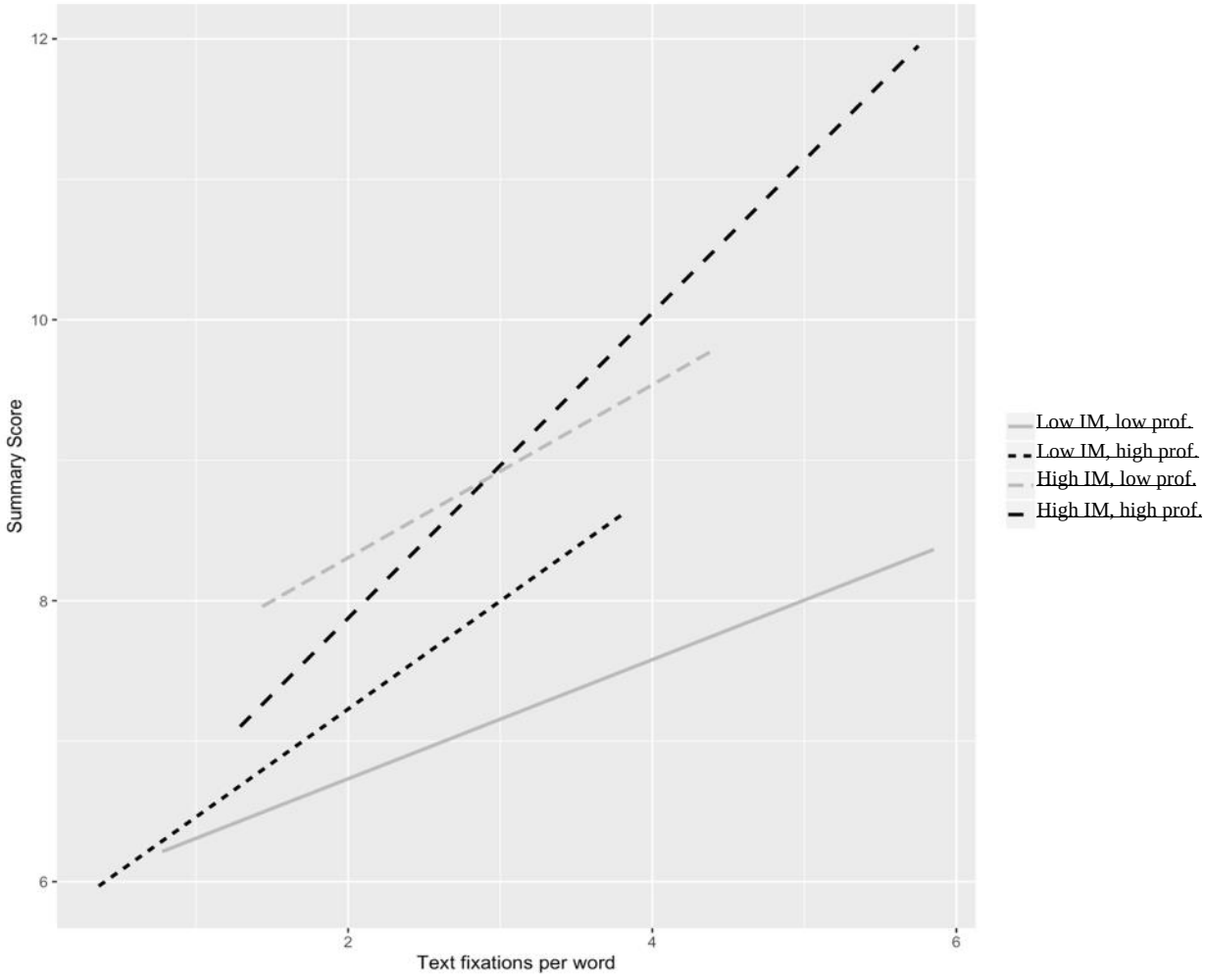


Figure 6.4.6 Summary score plotted against text fixations per word, with groupings for above-median and below-median motivation and Morpho-syntactic proficiency.
Note: IM = Intrinsic Motivation, prof. = Morpho-syntactic proficiency

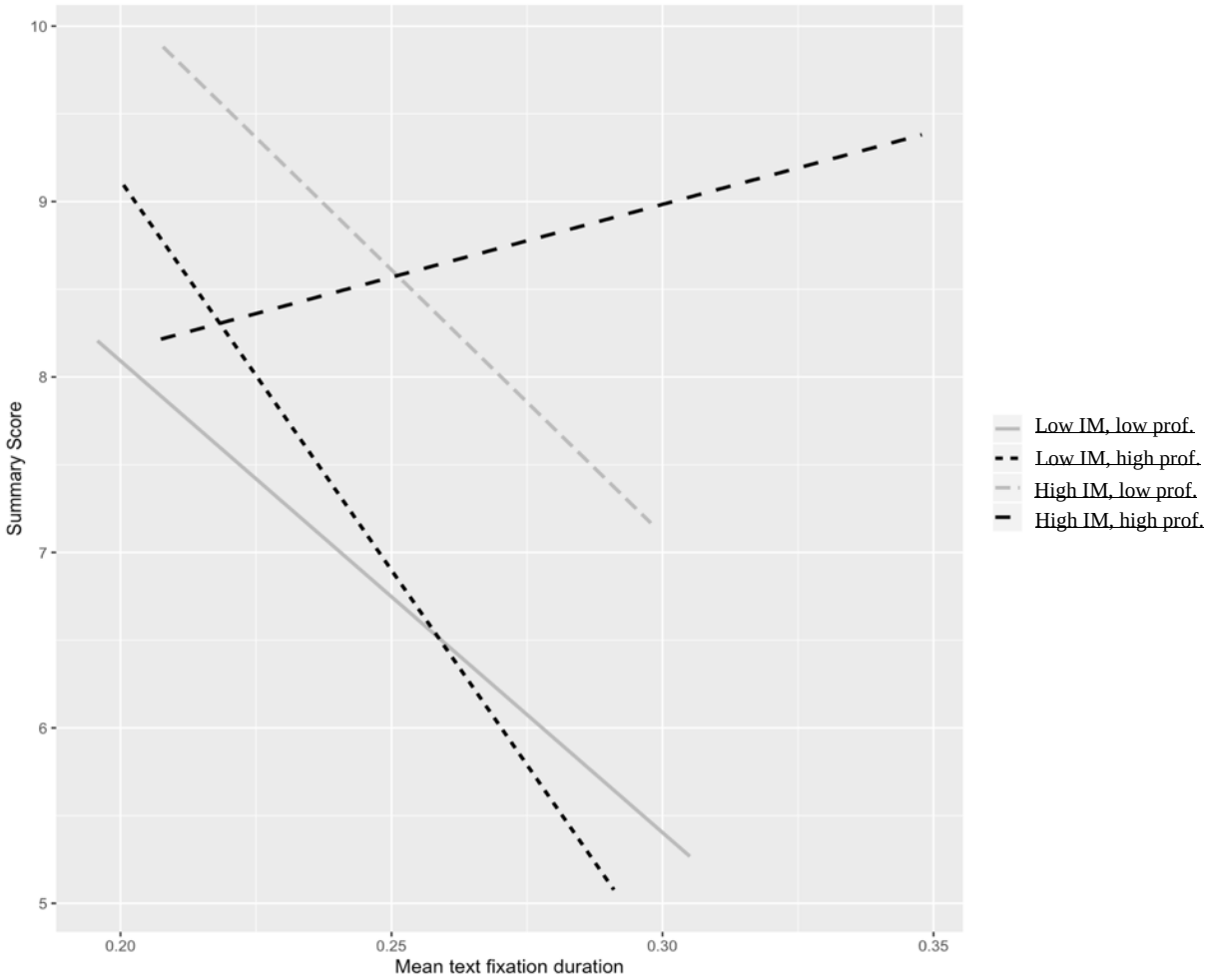


Figure 6.4.7 Summary score plotted against mean text fixation duration, with groupings for above-median and below-median motivation and Morpho-syntactic proficiency.
Note: IM = Intrinsic Motivation, prof. = Morpho-syntactic proficiency

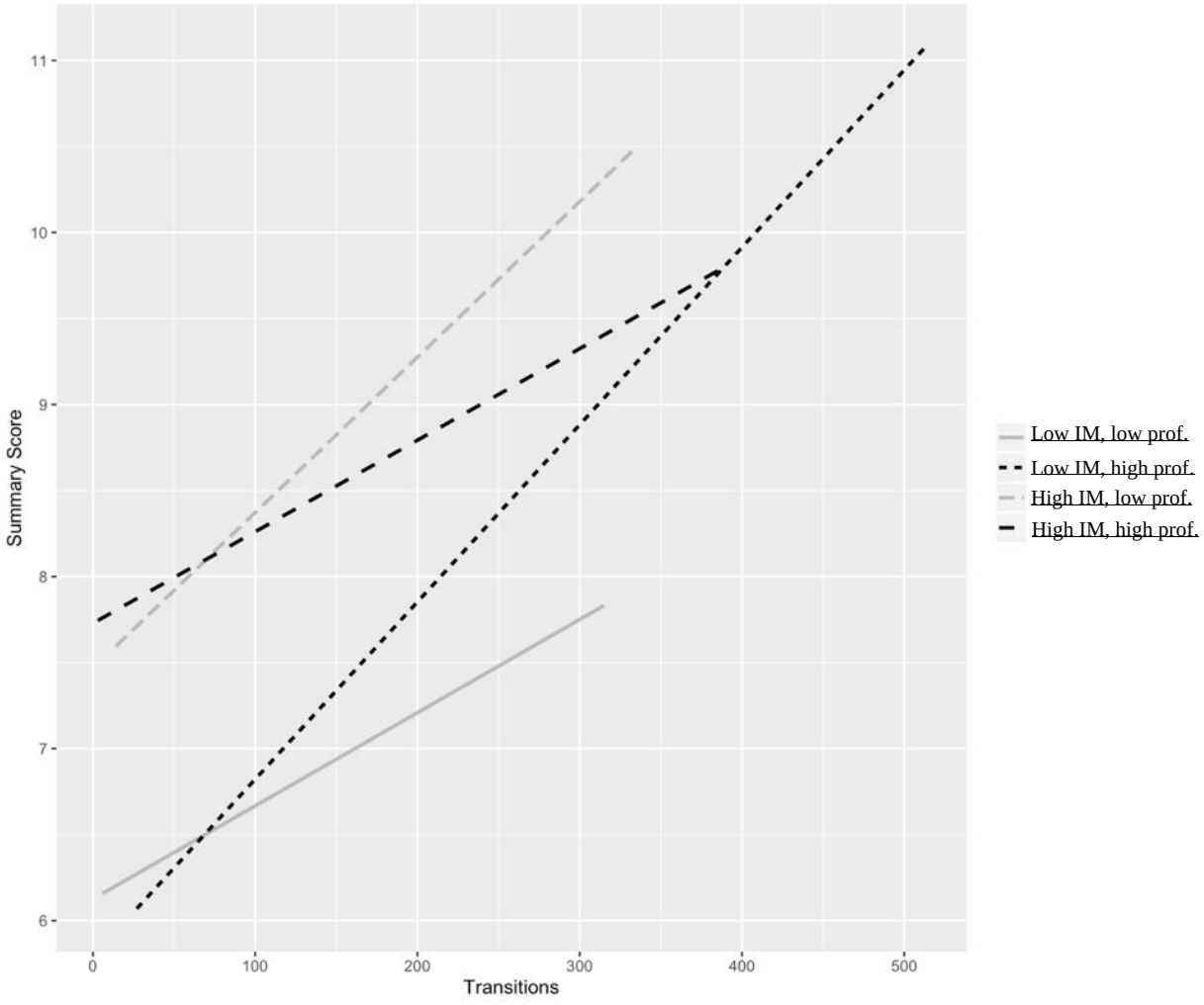


Figure 6.4.8 Summary score plotted against number of transitions, with groupings for above-median and below-median motivation and Morpho-syntactic proficiency.
Note: IM = Intrinsic Motivation, prof. = Morpho-syntactic proficiency

The three-way interaction with Morpho-syntactic proficiency, intrinsic motivation and mean fixation duration was found to be significant. In a similar fashion to the interaction effect found in the cloze score model, mean text fixation duration was negatively correlated with score, but for high motivation, high proficiency learners the trend was different (see Figure 6.4.7). Learners with both high motivation and high proficiency showed higher summary scores in general and higher summary scores even as mean text fixation duration increased. This was in contrast to learners with either low motivation or low proficiency who showed higher summary

scores as a result of shorter mean text fixation duration. The pairwise interaction between Morpho-syntactic proficiency and fixation duration was also significant, indicating that higher Morpho-syntactic proficiency mitigated the negative relationship between fixation duration and score at higher levels of Morpho-syntactic proficiency exists such that high proficiency learners mitigated the negative impact of slower mean text fixation duration on summary scores, while lower proficiency learners showed lower summary scores were related to greater mean text fixation duration. Among the main effects, text fixations per word was the most significant predictor, and no interactions with text fixations per word were significant. Independent of other predictors, higher number of fixations on text was a moderate positive predictor of higher summary scores. Higher motivation and shorter mean text fixation durations also had main effects in the model, predicting higher summary scores. Morpho-syntactic proficiency was predictive as a main effect in this model but was not more predictive than its interactions with mean text fixation duration.

Table 6.4.6 Linear regression model to predict summary task scores

Predictor	B	SE	*t*	*p*	r_2	Δr_2
Intercept	0.004	0.083	0.043	0.966		
Intrinsic Mot. x Morphosyntax x mean fix. duration	0.253	0.087	2.902	0.005*	0.058	
Morphosyntax x Mean fix. duration	0.211	0.078	2.707	0.008*	0.063	0.005
Intrinsic Motivation	0.188	0.086	2.188	0.031*	0.149	0.086
Morphosyntax	0.177	0.087	2.044	0.044*	0.207	0.058
Fixation per Word (Text)	0.385	0.086	4.498	< .001*	0.331	0.124
Mean Fixation Duration (Text)	-0.275	0.089	-3.098	0.003*	0.397	0.066

B = standardized coefficients, * significant at $p < .05$

6.5 Discussion

This final section in chapter 6 connects the results from the two studies. Connections are made between statistical analyses and visual data, and additional connections are made to previous research. Recommendations for L2 assessment and literacy development are offered. It ends with limitations of the study and future directions for research, of which there are many.

6.5.1 Summary and interpretation of findings

This chapter presented results to the second research question of this study, *to what extent does real-time reading behavior, as measured by eye-tracking, differ between reading tasks, and to what extent do online reading behaviors predict variance in reading comprehension scores beyond that predicted by individual differences?* Each part of this question was approached using real-time data from eye-tracking during participants' completion of reading tasks. Ten eye-tracking metrics which related to text-level reading and integration of reading material with non-text areas of interest (i.e. task areas). These were metrics related to global reading passage and comprehension task attention (fixations per word on text and on task, mean length of saccade, transitions between text and task, line and paragraph rereading), metrics related to careful/expeditious reading (mean length of fixation duration on text and on task), and metrics related to linearity of reading (mean number of fixations per dwell in text lines and text paragraphs). These measures were derived from raw fixation location, duration, and sequence collected from participants as they read three texts and completed three respective comprehension tasks (cloze, MC questions, and summary). Previous eye-tracking research suggests verification of statistical results with visual evidence (Kurzhals et al., 2017; Raschke et al., 2014). Thus, in interpreting these results, visual evidence from scan-paths and heat maps are referenced to provide extra explanation.

6.5.1.1 Research Question 2a

Eye-tracking measures were compared between the three tasks. Each task elicited different patterns across the eye-tracking metrics, allowing a few generalizations to be made across tasks. One is that fixations per word were greater for each task on task areas of interest than on text areas of interest, and this is reflected in the visual data (see Appendix G) which shows greater intensity of fixations on task areas than on portions of text. Additionally, fixations per word on text tended to be strongly related to rereading metrics in each task. This may indicate that the reading needed for comprehension in each task begets a level of rereading such that it is multicollinear with overall number of fixations. However, it can also be a limitation of the rereading measurement and other methods of recording rereading may trend differently from total fixations. Rereading in this study was measured as simply duration of second-pass dwells, or *look-backs*. Other measures of global regressive eye-movement exist, such as *look-froms*, i.e. the likelihood that a section of text induces a regression to earlier text, or *reinspections*, or the amount of regressive eye-movement done within a single dwell on an AOI.

There are noticeable differences in eye-tracking metrics between the three reading tasks. The cloze task involved the most fixations per word in the text and task AOIs, longest mean fixation duration on the text, largest number of mean fixations per paragraph dwell, the longest average rereading times by line dwell and by paragraph dwell. The MC task involved the largest mean number of fixations per line dwell and second largest mean number of fixations per paragraph dwell but was otherwise had the smallest measurement among the tasks for most of the metrics. The summary task involved the largest number of transitions, the longest mean length of saccade, and the longest mean fixation duration in both the text and task areas of interest.

The cloze task, in particular, involved the most fixations per word in both the text and task areas of interest, the longest average fixation duration on words in the reading passages, and the highest average fixations per paragraph dwell. Additionally, visual eye-tracking information showed that the cloze task involved attention to the entirety of the reading passage (although fixations become scarce in the cloze task after the final gap) and that the longest text fixations during the cloze task were clustered around the local reading context of cloze gaps with much less attention to other parts of the text (see Appendix G). This is indicative of careful reading at a local level. The cloze task thus involves activation of two goal-setting processes during reading: primarily careful, local decoding at and around gaps, and more expeditious reading farther away from gaps. This concurs with conclusions in previous research regarding the reliance on language proficiency and primarily local processes elicited by the cloze format (Kintsch & Yarbrough, 1982; Markham, 1985; O'Dell et al., 2000).

It is thus unclear whether the cloze task can be seen as a measure of higher-order reading comprehension, and it may be better suited as a general language proficiency task or measurement of lower-order reading processes. Other cloze formats may mitigate the emphasis on lower-order processes, either by providing word banks to mitigate vocabulary knowledge or more selectively targeting words which the test-taker is assumed to have topic knowledge about, though this may be difficult to achieve. The cloze tasks in this study were designed to target near-synonyms of already mentioned concepts, non-topic-dependent words, and cohesive links, each of which require text-schematic and top-down reading ability to process, yet this was not apparent in the results. Despite the reliability of the cloze test, it did not elicit global text comprehension as would be expected from a test of reading comprehension.

The summary task involved similar fixations per line dwell to the cloze task but fewer fixations per paragraph dwell. It also involved the greatest number of transitions between text and task, the longest average fixation duration in the task area, and the longest average length of saccade compared to the other tasks. Visual data (Appendix G) showed attention to the majority of the text but with specific regions of high-intensity repeated fixations. Inspection of visual data additionally verified that mean length of saccade and number of transitions were connected. The summary task was marked by many short dwells in paragraphs during summary writing, as participants briefly returned to paragraphs to find specific pieces of information before returning their gaze to the task area. Longer saccades in the summary task took place while participants were engaged in writing the summary, making frequent scans back to the text to identify information for use in their respective summaries. These returns to the text often involved multiple long saccades between paragraphs to reidentify the paragraph in which certain information was contained. The reading in the MC task and cloze task was more linear, and the saccades were typically between words in close proximity.

Summary tasks involved more attention to the entirety of the reading passage with greater intensity and a slower rate of reading than in the MC task. The summary task was also marked by longer average fixation durations in the summary writing area of interest. Together, this indicates the summary task elicited goal-setting strategies of both careful, global reading and selective expeditious scanning during writing. This interpretation is in line with previous conceptions of integrating reading and summary writing as eliciting higher order global and careful reading processes (W. P. Grabe & Stoller, 2013; Khalifa & Weir, 2009; L. Taylor, 2013). This match between global and selective reading for summary success was reported in previous research. For instance, Hyönä et al. (2002) found that L1 readers who paid relatively more

attention to topic marked areas of text were most successful in summary writing, and Prichard & Atkins (2019) found that selective attention to relevant sections was critical to success in L2 recall tasks. Compared to the cloze and MC tasks, summary tasks in this study appeared to more strongly activate higher-order reading processes which can be measured based on the reading behavior elicited.

The MC task elicited the fewest overall fixations (in text or task), shortest average fixation duration, the fewest transitions, and shortest mean length of saccade. It also involved the highest number of fixations per line dwell. Visual data from the multiple-choice tasks (Appendix G) shows few particular areas of fixation intensity and overall rather uniform attention to the texts. This indicates that MC tasks elicited linear, evenly distributed, expeditious reading behavior as one would expect to find for skimming. Although there were some long saccades for each participant in the MC tasks as they looked to the question area, these were far less frequent than in the summary task. Previous research has expressed concern that reading during MC tasks over-involves scanning and is akin to problem-solving rather than reading (Rupp et al., 2006), but the reading behavior associated with the MC task in this study was somewhat closer to linear skimming, which is similarly expeditious but not at the local level of scanning. The participants in this study were free to attempt questions at any point during reading, so some readers may have skimmed then answered, and some may have read questions and scanned. What is clear is that MC task elicits linear, expeditious text reading. Since few reading behaviors were positively associated with the MC task, it is difficult to assert what level of reading processes were activated by the MC task, but it is likely that the level of reading processes activated is between cloze and summary in terms of global text comprehension.

It has been well-understood in L1 and L2 reading assessment research that reading task will motivate different goal-setting processes related to reading rate and attention (Khalifa & Weir, 2009; O'Reilly et al., 2018b; Urquhart & Weir, 2014; Z. Wang et al., 2017), and the results of the logistic regression model constructed using eye-movement metrics adds further evidence of this. The model was found to be significantly more predictive of task than chance, indicating that the different reading comprehension tasks are motivating unique goal-setting processes in the reader. The model was very accurate in predicting summary tasks, and moderately accurate in predicting cloze tasks, but had more difficulty classifying MC tasks, more often than not classifying them as summary tasks. This likely stems from the differences between most of the included variables with the cloze task, and the paucity of predictors that could reliably distinguish MC and summary tasks. There is also inherent difficulty in predicting between more than two categories using this method. Logistic regressions assign a category based on where a single numeric result falls along a single dimension. However, the tasks in this study aligned differently for the various reading behaviors. For example, the MC task elicited the highest mean fixation per line dwell and the cloze task elicited the lowest, and the summary task was somewhere in between. Yet the summary task was not always the middle category, since the summary task elicited the longest mean length of saccade and largest number of transitions. Thus, the middle category for prediction was bound to be tenuous.

In the model, the strongest predictor of task based on log odds was the number of transitions between text and task, with higher numbers of transitions favoring the summary task. Fixations per word in text and task were the next strongest predictors and were associated with the cloze task. Fixation per dwell measures also contributed significantly to the model, with fixations per paragraph dwell slightly favoring cloze scores, and fixations per line dwell slightly

favoring summary tasks. Fixations per dwell have not been extensively researched in applied linguistics and language assessments, but there is evidence from psychological research that more fixations per duration relate to heavy cognitive demand and decision making (Klichowicz et al., 2016). Although the cloze task was more associated with features signifying careful local reading, that larger mean fixations per paragraph dwell predicted cloze tasks, showing that it is still a cognitively demanding task, but may not demand higher-order reading processes.

It is also worth noting that the model was likely to mis-classify reading for MC questions as reading for summarizing. This may come from the positive predictive odds of the mean fixation per line dwell, which was the metric most positively associated with the MC task. This indicates that the two tasks (MS and summarizing) may be more similar to each other than the cloze task, a fact obviated by the physical design of the tasks. Both tasks involved reading a text in parallel to a separate task area, unlike the cloze task which had the task "area" within the text itself. The task layout for MC and summary entails that the basic logistics of reading behavior, needing to process text and then transition to a distant task pane to provide response, is similar for these parallel tasks. However, the level of careful reading and attention is different for both.

6.5.1.2 Research Question 2b

Section 6.4 of this chapter presented results for the second part of research question 2: *to what extent do online reading behaviors predict variance in reading comprehension scores beyond that predicted by individual differences*? The correlation data shows that performance on each task was related to a different set of eye-tracking metrics measuring reading comprehension behavior. The eye-movement behavior elicited by the three tasks was not necessarily the type of behavior which was most conducive to better performance. Additionally, the set of significant predictors of score from eye-tracking metrics were unique to each task, although some

similarities existed. In each task, fixation duration in both text and task areas was negatively correlated with score, though this was not always significant. This is attested in previous research which shows that more skilled readers typically make shorter, more efficient fixations (Ashby et al., 2005; Bax, 2013; Krieber et al., 2016). Although each task elicited more fixations per word in task areas than text areas, fixations per word on task areas of interest was negatively correlated with score in each task

Performance on the cloze task was correlated negatively with transitions, fixations on the cloze gaps (task areas), and fixation duration on the reading text. Additionally, fixations per word on task and transitions were multicollinear with overall text fixations per word. This is unsurprising given the nature of the cloze task, where task areas are in-line with the text. Taken together, these relationships show that higher performance on the cloze task was related to efficient reading and handling of the individual cloze gaps. Of these measures, only mean fixation durations on the text area contributed significantly to the model predicting cloze score. Although mean text fixation duration was negatively correlated with cloze score alone, there was a significant interaction effect between mean text fixation duration, reasoning, and Morpho-syntactic proficiency. As was seen in the interaction graph (Figure 6.4.1), at higher levels of both proficiency and reasoning, longer fixation durations became predictive of higher cloze score.

This effect indicates that as reasoning and proficiency increase, the effect of longer fixation durations is a positive predictor of score. This three-way interaction is difficult to interpret, but this could indicate either that careful reading is more important than efficient reading for more proficiency and logical readers, or that there is diminishing returns regarding local processing efficiency's impact on cloze performance. Though longer fixation durations predicted higher scores at higher levels of Morpho-syntactic proficiency and reasoning, its main

effect is negatively predictive of cloze score for test-takers at lower levels of Morpho-syntactic proficiency, reasoning, or both. This indicates that the interaction effect represents the compensatory function Morpho-syntactic proficiency and reasoning may have for readers with slower processing evidenced by long average fixation duration. Nevertheless, the main effect of high Morpho-syntactic proficiency was the strongest predictor of higher performance on the cloze task, indicating that given average levels of reading speed and text fixation duration, Morpho-syntactic proficiency remained a strong influence on cloze performance. This underscores the importance of lexico-syntactic knowledge on cloze score.

Higher performance on the MC task was correlated negatively with transitions, task area fixations per word, and question area fixations duration. These relationships indicate that the reading behavior which related to higher MC scores was not connected to text reading behavior, which was overall linear and expeditious, but rather to the efficiency with which readers attended to the questions. The metrics associated with the MC task in the between tasks comparison, such as fixation per line dwell (see section 6.2, this chapter), were not strongly associated with score. Rather, two task area-related metrics not associated with the MC task, transitions and mean fixation duration on task areas, were predictive, and both were negatively correlated. The model to predict MC score showed that about 20% of variance in MC scores could be accounted for by text-to-task transitions and fixation duration on the questions, with higher scoring participants making fewer transitions between text and questions and shorter fixation durations on the questions.

The fact that MC score correlated with reasoning ability and was modeled by behaviors related to efficient reading of questions and answers indicates that readers likely used logic and test-wiseness strategies at least as much as text comprehension to complete the MC task. This

may relate to the previous assertion that success on MC tasks relates to efficient problem solving skills (Rupp et al., 2006). Visual inspection of MC task eye-movement patterns (Appendix G) also indicate that the MC task encouraged primarily attention to the questions compared to attention to the text, and that attention to the text was rather uniform across paragraphs, but selective within paragraphs, with few strong points of attention clustered in particular lines. This is similar to findings regarding MC choice question responding in O'Reilly et al., (2018), where participants read sections rather linearly once the question-relevant segment of text was identified. In sum, it is likely that the readers' interpretations of questions, more-so than their processing of text, influenced MC task performance. This indicates that success on the MC task was not dependent on careful global text processing, or else was perceived as easy enough by test-takers for them to not rely on top-down processing. Since, the model to predict MC scores was weak by comparison to the other tasks, there is still much variance in MC scores unaccounted for, and it may relate to an unmeasured latent, efficient reading construct. A possible advantage of the MC task is the mitigation of proficiency, at least with the advanced academic readers who participated in this study. The MC task scores were less correlated with L2 proficiency than the other task scores, which may indicate the strength of the MC task is eliciting expeditious reading skills while mitigating language production ability (Genesee & Upshur, 1996b).

Score on the summary task showed a markedly different pattern of correlation than the other two comprehension scores. Unlike in the cloze and MC tasks, transitions between text and summary writing area were found to be positively correlated with a moderate effect size, and fixations per word in the reading passage was additionally found to be positively correlated with a moderate effect size. Similar to the other tasks, summary score was negatively, albeit weakly,

correlated with mean fixation duration on the reading text. This indicates that higher summary performance is correlated with efficient reading (shorter fixations), but more extensive coverage of the text (more fixations), and that higher quality summary writing was related to more return looks at portions of text during writing. The metrics associated with the summary task in the between tasks comparison, mean length of saccade, transitions, and task area fixation duration (see section 6.2, this chapter), were not strongly associated with summary score. Rather, two text reading-related metrics were associated with the summary task: fixations per word on text and mean fixation duration on text.

In the summary score model, similar to the cloze score model, mean fixation duration exhibited an interaction with individual differences in the model. For most test-takers, mean text fixation per word was a negative predictor of score, i.e. shorter fixations are better for performance. However, at higher levels of Morpho-syntactic proficiency and intrinsic reading motivation, longer mean fixation duration was more predictive of summary score. An interaction between just mean text fixation duration and Morpho-syntactic proficiency effect was also predictive, indicating at average motivation levels, the change in effect direction of fixation duration across Morpho-syntactic proficiency levels remained. As before, this could indicate that at higher proficiency and motivation levels, there are diminishing returns for the efficient processing in making shorter fixations, or that there is a compensatory effect of Morpho-syntactic proficiency and motivation for slower processors. The latter hypothesis may be more tenable, given the smaller main effect for Morpho-syntactic proficiency in the model compared to the total effect of the interactions. The summary task may impose greater linguistic and motivational demands on readers who are less efficient text processers. As a main effect, a larger number of fixations per word in text was significant independent of individual differences,

explaining 12.4% of summary score variance. Motivation also had a significant positive main effect on summary score, as did Morpho-syntactic proficiency, though these main effects were weak. Mean fixation duration had a significant negative main effect in the model, reflecting its overall negative correlation with score.

The connection between higher numbers of text fixations per word and success on the summary indicates that there are instances where better performance is associated with more careful, perhaps less efficient, reading. This may indicate that the summary task pushes readers to build the most intricate mental model (Bax, 2013), but it may also indicate that the task in general demanded more in terms of cognitive load and perceived difficulty. Taken together, higher performance on the summary task required greater fixations per word in the text, indicating it necessitated more careful global reading, and that readers with some combination of higher motivation, shorter mean fixation duration, and higher proficiency performed better. In summary, this indicates that success on the summary task was more dependent on reading behavior expected for careful, global reading for the purpose of higher-order comprehension, although it may be a cognitively demanding and perceptibly difficult task.

6.5.2 Conclusion, limitations and future directions

This study is unique in that rather than comparing eye-tracking measures between participant groups (e.g. high and low skilled readers) or measuring eye-tracking in relation to specific lexical and syntactic features, this study compares eye-tracking metrics between difference reading tasks and task performance. Two implications for testing can be discerned from the above results. First, the differences between tasks in the types of reading elicited indicates that a variety of reading comprehension tasks at various levels of cognitive involvement are necessary to cover the different types of reading. Additionally, for learners and

educators, it is worth reinforcing the importance of goal-setting strategies as part of successful reading. Since efficient reading was associated with score in the discrete tasks, but global reading was associated with score in the open-ended summary task, developing readers should develop awareness of reading purpose and tailor their reading speed to the demands of reading purpose.

There are several limitations in the current study. First, the eye-tracking measures used in this study are quite coarse. Fixations per word and average fixation duration, for example, are very general measurements based on a participant's entire reading trial worth of data. The areas of interest in this study were coarsely defined to understand whether readers were paying attention to the text or to the task in the reading trials. However, a more principled selection of areas of interest may also provide illustration of reading behavior across different reading settings. There is plenty of room to investigate more finer grained eye-tracking metrics at specific paragraph, sentence, and word levels. There was also no examination of how the eye-tracking metrics varied within participants over the time course of trials. Since rereading measurements were not statistically distinct from total number of fixations per word in this study's data, examining rereading by examining eye-movements at different times throughout trials may provide better insight to the conscious strategies of readers, such as when and where to reread text. Finally, no linguistic features were highlighted as areas of interest, and the current study took a rather content-agnostic approach to eye-movements in the hopes that task conditions rather than topic information and linguistic features could be witnessed as motivating reading behavior. However, development of areas of interest based on a comparison of task response areas to related text information could provide further insight into how eye-movements relate to accessing and processing specific information from text.

Next, some eye-tracking metrics used in this study are a matter of individual differences and general literacy, so it may be difficult to make a claim that the reading task elicited a certain behavior or that a skilled reader consciously activated use of behavior for the task. It is difficult to make claims about whether shorter fixations lead to better reading scores, although it was predictive in each model, as it is not clear whether being a skilled reading causes one to make shorter fixations, or making shorter fixations helps one develop into a skilled reader. Understanding this would require further investigation. However, the idea that the eye-tracking metrics are mere individual difference factors is mitigated by the within individual comparison of the between tasks analyses. For the task comparisons, eye-tracking metrics were compared based on how they differed within a single reader across tasks, so the difference between fixation duration between tasks retains interpretability.

Finally, previous research warns against claiming that any eye-tracking measure is direct evidence of certain underlying processes (Cook & Wei, 2019). To address this, findings were discussed in terms of the intersection between eye-tracking features which related to tasks and performance. The conglomeration of metrics associated with each task allow for some inference of underlying process, but the connection between metrics and cognition, such as fixation duration and careful attention, should be taken with a grain of salt.

In addition to the adjustments and additions to eye-tracking metrics mentioned above, there are several avenues for further research. Previous research has looked at how eye-tracking can be used to understand processes in answering shorter open-ended response comprehension questions (Bax, 2013). The current study examined eye-movement behavior in MC tasks, cloze tasks, and summary tasks, but clearly there is a larger gap in openness and productivity between the summary task and the MC and cloze tasks. A task with more open-endedness than the cloze

task, but not as productive as the summary task, may have provided more insight into how tasks elicit reading behavior, and comparing short answer tasks with the other tasks should be investigated in future research.

The productive nature of the summaries are of particular interest for future research. Linguistic features of reader production can be analyzed using natural language processing methods and potentially compared to eye-tracking data to explore the relationships between attention and language production, perhaps providing insight into the processes of developing mental models. Additionally, qualitative examination of visual data from eye-tracking was only briefly utilized in this study, but there is room to explore further the visual data from heat maps and scan paths as they allow us to witness real-time strategy use. Future studies can examine the appearance of reading strategies in visual data.

Last, motivated by the importance of intrinsic motivation in the summary score, it is important to understand how individual readers' motivation may affect their eye-movement behavior during reading. Reader perceptions, as gathered by stimulated recall, interview, or survey, may provide further cues to aspects of readers which impact the way they approach texts. Further studies should include self-reported data from participants regarding perceptions of topic familiarity, task ease, and test authenticity which can be compared to the actual real-time reading behavior of readers.

7 CONCLUSIONS

This chapter presents a general discussion of this dissertation. It includes a summary of the research carried out, a synthesis of the findings for each research question in this dissertation, further connections to previous research, and more in-depth recommendations for language testing and education.

7.1 Answers to research questions

From the results in chapter 5, several conclusions can be drawn. Regarding the first part of research question 1, whether examinees respond significantly faster to sentences inferable from a text than to unrelated sentences after reading a text, the results from this study show that related sentences are responded to significantly faster than other types of sentences in a sentence verification task following passage reading. This difference in response speed is not dependent on the type of reading task completed during passage reading. Regarding the second part of research question 1, the extent to which inference generation predicts variance in comprehension task outcomes (scores) independent of proficiency and individual differences, the results from this study show that A) inference generation only influences reading outcomes when the measured reading score is explicitly designed around an aspect of reading where inferencing is critical (i.e. mental modeling in the summary task) and B) the impact of inferencing on scores is secondary to that of Morpho-syntactic proficiency and intrinsic motivation when it is predictive of scores.

Chapter 6 presented findings from analyses of eye-tracking metrics measured during online reading comprehension task completion. This study is unique in that rather than comparing eye-tracking measures between participant groups (e.g. high and low skilled readers) or measuring eye-tracking in relation to specific lexical and syntactic features, this study

compares eye-tracking metrics between different reading tasks and task performance. In response to the first part of research question 2, eye-tracking metrics were able to distinguish reading during each of the three test tasks. Reading during MC tasks was marked by more fixations per line dwell, shorter fixation durations, fewer overall fixations, shorter average saccades, and fewer transitions between text and task. Reading during the cloze tasks was marked by more overall fixations per word, longer mean fixation durations on the reading text, and more fixations per paragraph dwell. Reading during the summary task was marked by longer fixation durations on the task area compared to the other tasks, longer average saccades, and more transitions between text and task. The tasks elicited different reading patterns, and the reading patterns related to higher scores on the tasks also differed.

Regarding the second part of question 2, eye-tracking metrics contributed predictive power to models of scores in each comprehension task. On the MC task, score was related to some of the reading behaviors already associated with the MC task. Higher scores were predicted by shorter mean fixation duration on questions, fewer fixations per word on the questions, and fewer transitions between text and questions. The former two metrics were predictive of score, indicating more efficient attention to the questions predicted MC score. In this way, success on the MC task was a matter of less is more.

Although the reading behavior the cloze task elicited involved more fixations on the text and cloze blanks and longer mean text fixations, cloze score was negatively correlated with duration of fixation and attention to cloze blanks in terms of transitions and fixations per blank. This is consistent with previous research which found that efficient fixation is related to comprehension (Bax, 2013; Rayner et al., 2006). Unlike in the case of the MC task, where the behavior elicited by the task was also conducive to higher scores, the behavior associated with

cloze tasks, e.g. longer fixation durations on the text, were not beneficial to higher scores. Despite the negative correlations with fixation duration, in the full model predicting cloze scores, there was a positive interaction between Morpho-syntactic proficiency, reasoning, and fixation duration, indicating that higher levels of Morpho-syntactic proficiency and reasoning could offset the negative impact of making longer fixations on cloze score. A main effect for shorter fixations was also predictor of higher scores, meaning that at mean Morpho-syntactic proficiency and reasoning scores (or lower), processing efficiency was an important predictor of higher cloze score. Positive main effects on cloze score also were found for higher proficiency and reasoning.

So far, the models predicting scores have showed that with eye-movement behavior during text reading, less is more. Conversely, summary scores showed positive correlations with text-to-task transitions as well as number of text fixations, but still showed a negative correlation with text fixation duration. Similar to the cloze model, in the predictive model of summary scores, there was an interaction between motivation, Morpho-syntactic proficiency, and mean fixation duration. The interaction effect was positive on summary score, indicating that as any of the three factors increase, the positive impact of the other factors increases. For Morpho-syntactic proficiency and motivation, which also had positive main effects in the summary score model, this showed that these individual differences can reinforce their impact on summary performance. For text fixation duration, which alone had a negative main effect on summary scores, the positive interaction indicates that increases in motivation and/or Morpho-syntactic proficiency can mitigate the negative impact of slower processing. An additional predictor of higher summary scores was higher numbers of text fixations, which was a moderate predictor of higher score independent of other variables. This shows that summary writing is benefited by a

combination of proficiency, motivation, and efficient text processing, but also predicted by global text attention.

7.2 General discussion

Various individual differences which relate to performance on reading comprehension tasks were analyzed in this study. Morpho-syntactic proficiency was found to play a role in both cloze task and summary task performance, and this relationship was much stronger in the cloze task. Reasoning ability, as measured by a non-verbal series completion test (see chapters 3 and 4), was correlated with cloze and MC task performance, but not summary quality. Instead intrinsic motivation, based on a survey, was a significant predictor of summary task performance. Working memory was not found to contribute to reading comprehension performance, and this was perhaps due to its non-independence from reasoning ability, which was often the stronger correlate of comprehension. Reading speed was surprisingly not correlated with any comprehension scores, and this may speak to the importance of goal setting in measuring reading. The reading speed task was rather purposeless from the perspective of participants in this study, who merely read a text and indicated when they were finished. The speed at which one reads simply to be done with a text may not be reflective of the reading speed in the more realistic assessment tasks used in the main study procedure.

Regarding the generation of inferences during second language reading assessment, the findings in this study show that inference generation did not occur to a more or lesser degree across the three tasks, indicating that inference generation is a component of advanced academic reading of English as an additional language regardless of task format. This is consistent with the position that inferencing is not always a conscious strategy, but some inferencing may instead be automatic as needed during reading (Cain & Oakhill, 2001) The inference generation measured

in this study did vary among individuals, and the degree to which readers activated inferencing while reading related to performance on the summary task. This indicates that tasks which push readers to construct a more detailed mental model of a text rely more on the level of a reader's inference generation. This provides evidence of the cognitive validity of the summary task for tapping into higher-order skills which are critical in academic reading.

The online reading behaviors measured in this study were used to distinguish the three test tasks, and the results provide evidence for the type of reading readers engage in when given specific reading tasks. The eye-movement behavior elicited by the cloze task present a profile of careful, local reading. This type of reading is related to lower-order decoding processes. Score on the cloze was related to proficiency, reasoning, and efficiency of fixations.

The eye-movement behavior elicited by the MC task were in line with expeditious reading, as evidenced by fewer average fixations per word and shorter average fixation duration, and linear reading, as evidenced by the higher average fixations per line dwell and shorter average length of saccade. This type of reading is in line with expeditious comprehension processes such as skimming (Urquhart & Weir, 2014), or else the efficient comprehension that occurs when processing a text perceived as easy (Grabe, 2009; Wallot, 2011). Perhaps to the task's credit, MC score was the only task to not be predicted by proficiency. However, the explanatory power of the predictive model was weak, and only eye-movement efficiency during question reading impacted MC score. In the absence of other predictive factors, the impact of fast question processing on score presents potential concern that MC tasks are overly susceptible to test-wiseness strategies.

The eye-movement behavior in the summary task was global and careful, indicated by number of fixations and fixation duration on text, but also included a degree of searching and

scanning, indicated by fewer fixations per dwell, longer saccades, and more transitions between text and task. This is the reading one would expect during careful text modeling and reading-to-learn (Urquhart & Alderson, 1984). Score on the summary task was predicted by motivation, proficiency, inference generation, processing efficiency, and the number of text fixations per word during reading. The contribution of this diverse set of variables related to some lower but mostly higher order reading processes speaks to the utility of the summary task.

7.3 Implications for assessment practice and instruction

Several implications for language testing can be discerned from the above results. For test design, the findings of this study provide evidence that higher-order reading skills can be captured in L2 reading tests if desirable, but the aspect of mental modeling must be explicitly built into design and scoring of the test. Each reading test task examined in this chapter had significant predictive models with unique sets of predictors. Thus, it is important for reading tests to utilize a variety of tasks to account for the many subskills which contribute to academic second language reading.

Depending on what one intends to measure by assessing reading in a second language, the findings of this study provide some guidance to the appropriate task. If one views reading as an extension of L2 proficiency, then the cloze task captures primarily lexico-grammatical proficiency and decoding ability. If removing the influence of L2 proficiency for advanced readers and measuring efficient, expedient reading ability is the goal of assessment, then the MC task can work to this degree. However, if one views second language reading as a complex mix of language proficiency and literacy strategy factors, and that a primary goal of a second language reading comprehension test is to capture a higher-order cognitive reading skill such as

inferencing and ensure that score is related to global text understanding, then the summary task with clear guidelines for evaluation based on mental modeling is most likely to capture this skill.

The differences between tasks in the types of reading elicited indicates that a variety of reading comprehension tasks at various levels of cognitive involvement are necessary to cover the different types of reading. The case can be made that the different tasks investigated in this study can be directed at different levels of reading and language ability, with MC questions more useful for lower-level learners, cloze tasks more useful for slightly higher-level learners, and summaries being better suited for learners at more advanced academic levels. The summary task was the most reliable task and was the task which involved the most complex modeling. Considering the population included successful advanced academic readers, this shows that the summary task may be best suited of the three tasks for assessing comprehension for this population.

Additionally, for learners and educators, it is worth reinforcing the importance of goal-setting strategies as part of successful reading. Since efficient reading was associated with score in the discrete tasks, but global reading was associated with score in the open-ended summary task, developing readers should develop awareness of the ultimate goal of comprehending a text, i.e. their reading purpose, and tailor their reading speed to the demands of the goal.

Returning to models of second language reading, the analyses in this dissertation indicate that there are factors which impact L2 reading comprehension beyond proficiency and individual differences associated with L1 literacy (reasoning and motivation). The fact that shorter fixation durations were related to higher cloze and summary scores, and interacted with other individual differences, indicates that efficient processing is a critical aspect of reading comprehension, and it cannot be strictly attributed to L2 proficiency or L1 literacy. Processing efficiency has instead

been traced to exposure to printed material from the language in question (Chateau & Jared, 2000; W. Grabe, 2010; Yamashita, 2008). For reading instruction, this provides evidence for the usefulness of task-oriented, extensive reading; building up learners' exposure to second language texts can improve processing ability. Extensive reading typically focuses on narrative texts, so it may be beneficial for teachers to incorporate more expository texts with specific reading goals into extensive reading programs. The exact relationship between extensive reading on shortening fixation duration during text reading has yet to be investigated, however.

7.4 Limitations and considerations for future research

There are several areas of limitation in this study, and subsequently many avenues for further research. The sample for this study included only matriculated university undergraduate and graduate students. Therefore, the ability to extrapolate results from this study to a general English language test-taker population is limited, as the sampled participants represent a group who have already proven themselves to be successful test takers.

The study highlights the difficulty in examining inferencing in expository texts, which are more information dense, put more responsibility on the reader to interpret information, are not necessarily linear, and require more specific background information for comprehension when compared to narratives, the type of text usually employed to understand inference generation (Lorch, 2015). In L1 reading literature, reading expository texts has been found to more likely trigger literal comprehension processes and discourage unnecessary inferencing past those necessary for local coherence (Noordman et al., 1992). Noordman et al. (1992) further assert that inferencing during expository text may be dependent upon goal setting, a conclusion for which the current study provides some support. For more precise understanding of

inferencing in academic L2 reading, further research is needed in general on inference generation while reading expository texts.

The current study also makes no practical distinction between the various types of inferences which could be made during reading, such as bridging inferences, causal inferences, or elaborative inferences, instead treating inferencing as a general ability to insert default or logical information into comprehension gaps. Although there is support for examining inferencing as a general skill (Kendeou, 2015), further research may examine specific types of inferences which can be drawn from expository texts to understand if specific types of inferencing contribute to comprehension.

Although data for many measures related to reading and language ability were collected for this study, one type of data which could not be collected was direct L1 literacy data. Due to the diverse pool of participants, an L1 literacy measure was unfeasible. By examining correlates of literacy, such as reasoning, working memory, and motivation, it was hoped that skills which may contribute to successful L1 literacy could be captured, but this is not guaranteed. As previous models of L2 reading generalize the components of L2 reading to be either L2 proficiency- or L1 literacy-based (Koda, 1988), it is difficult to situate the results of these findings. L2 proficiency was certainly found to be related to reading comprehension, more-so than other individual differences, but a comparison between L2 proficiency and a general literacy ability could not be compared here. Future studies may include L1 reading comprehension tests to create a fuller picture of the skills which contribute to L2 reading comprehension.

Reading speed and typing speed were measured using a simple text reading exercise and minute-long typing test respectively, but neither offered readers much of a purpose for reading the text or typing. The superficial nature of these tasks may have led to reading and typing speed

scores which were not reflective of realistic reading and writing demands, since reading speed and typing speed were not found to relate to comprehension score or figure significantly into predictive models. Future studies should ensure that reading speed tasks encourage authentic reading behavior in order to be an accurate measure and include a more interpretable measure of L2 writing proficiency beyond typing speed.

Another theoretical aspect glossed over in this study is the importance of textual features and difficulty on reading comprehension. On the one hand, texts in this study were selected for their similar nature and source (academic textbooks). Although text features were measured for control purposes, syntactic and lexical features play an important role in text processing and comprehension (Crossley, Greenfield, & McNamara, 2008; Crossley et al., 2017) and they may provide further evidence to behavior during comprehension. An unanswered question from this study regards whether reading behavior associated with higher cognitive demands are activated strategically by readers or activated due to processing difficulty and text complexity. Future studies should take into account the role of textual features and perceived text difficulty on reading behavior. Since the connection between reading comprehension and processing behavior was shown in this study, this also opens the door for studies which use eye-tracking metrics related to good and poor comprehension to understand text difficulty.

Regarding the comprehension test tasks, interpretation of the score modeling results is limited by the varying reliability of the tasks. For the summary and cloze tasks, the reliability is high enough to warrant generalizations about the task based on the models in chapters 5 and 6. However, the reliability of the short MC tasks was lower overall, and this impacts the ability to interpret the score models. It may be the case that with more items at more varied difficulty

levels and higher reliability, the outcomes of linear models to predict MC score would have different outcomes.

Many alternatives to the three tasks presented in this study are worth further research. Although the MC task, cloze task, and summary task each had specific reasons for inclusion, results may differ in the cross-task comparisons if other tasks were included. Other forms of selected response task types, banked gap-fill task, and short-answer constructed response tasks are each worthy of further analysis and are not reflected in the results in this study.

From a methodological standpoint, there are several limitations. Although over 100 participants were recruited for this study, this is still a relatively small sample size considering the types of analyses conducted, especially after accounting for outliers and missing data. The linear models to predict reading task scores may suffer from low power. Post-hoc power analyses based on real effect sizes found the average power of the linear models in this study to be around 61% on average, less than anticipated in the a priori analysis. Thus, the chance for false negatives are fairly high, and future studies taking a similar approach to understanding inferences in L2 reading will require larger sample sizes.

Another quantitative limitation was the number of sentences supported by the SVT. As the expository texts used in this study were fairly short, information dense, and targeted toward those with little background knowledge on a given subject, there were few opportunities to isolate inferable ideas from the texts, and thus few data points to rely on for each sentence condition for each text. Future studies can utilize this method with a larger pool of source texts of various lengths and a larger pool of related sentences.

Additionally, regarding the use of SVT in this context, previous uses of SVTs were typically used to understand differences between types of stimuli and experimental conditions. In

this regard, the current study contains findings of this type, with sentences related to the priming reading passage responded to faster than unrelated sentences. However, there is less use of SVT to understand within person differences, and the task may be less suited for this purpose. It is thus unsurprising that the relationship between the post-hoc SVT and reading comprehension performance was weak overall.

Regarding limitations within the eye-tracking analyses, the eye-tracking measures used in this study are quite coarse. Fixations per word and average fixation duration, for example, are very general measurements based on a participant's entire reading trial worth of data. The areas of interest in this study were coarsely defined to understand whether readers were paying attention to the text or to the task in the reading trials. However, a more principled selection of areas of interest may also provide illustration of reading behavior across different reading settings. There is plenty of room to investigate more finer grained eye-tracking metrics at specific paragraph, sentence, and word levels. The distinction between task and text AOIs differed quite drastically between the cloze tasks and the other tasks, although this did not seem to firmly distinguish the cloze task form the others in the analyses, as the cloze task was closer to the mean in terms of fixation duration on task and task to text transitions, so this may not have been as much of a liability as it would appear on the surface.

No examination of how the eye-tracking metrics varied within participants over the time course of trials was conducted, and time itself was not included as a factor. Although participants were cut-off after 20 minutes, their relative time expenditures between the tasks were different, with the summary task typically taking more time and including longer gaze durations on text and task (before controlling for words). Total time on task may be an important factor that should be controlled for or included in future analyses.

Since rereading measurements were not statistically distinct from total number of fixations per word in this study's data, examining rereading by examining eye-movements at different times throughout trials may provide better insight to the conscious strategies of readers, such as when and where to reread text. No linguistic features were highlighted as areas of interest, and the current study took a rather content-agnostic approach to eye-movements in the hopes that task conditions rather than topic information and linguistic features could be witnessed as motivating reading behavior. However, development of areas of interest based on a comparison of task response areas to related text information could provide further insight into how eye-movements relate to accessing and processing specific information from text.

Next, some eye-tracking metrics used in this study are a matter of individual differences and general literacy, so it may be difficult to make a claim that the reading task elicited a certain behavior or that a skilled reader consciously activated use of behavior for the task. It is difficult to make claims about whether shorter fixations lead to better reading scores, although it was predictive in each model, as it is not clear whether being a skilled reading causes one to make shorter fixations, or making shorter fixations helps one develop into a skilled reader. Understanding this would require further investigation. However, the idea that the eye-tracking metrics are mere individual difference factors is mitigated by the within individual comparison of the between tasks analyses. For the task comparisons, eye-tracking metrics were compared based on how they differed within a single reader across tasks, so the difference between fixation duration between tasks retains interpretability.

Finally, previous research warns against claiming that any eye-tracking measure is direct evidence of certain underlying processes (Cook & Wei, 2019). To address this, findings were discussed in terms of the intersection between eye-tracking features which related to tasks and

performance. The conglomeration of metrics associated with each task allow for some inference of underlying process, but the connection between metrics and cognition, such as fixation duration and careful attention, should be taken with a grain of salt.

In addition to the adjustments and additions to eye-tracking metrics mentioned above, there are several avenues for further research. Previous research has looked at how eye-tracking can be used to understand processes in answering shorter open-ended response comprehension questions (Bax, 2013). The current study examined eye-movement behavior in MC tasks, cloze tasks, and summary tasks, but clearly there is a larger gap in openness and productivity between the summary task and the MC and cloze tasks. A task with more open-endedness than the cloze task, but not as productive as the summary task, may have provided more insight into how tasks elicit reading behavior, and comparing short answer tasks with the other tasks should be investigated in future research.

The productive nature of the summaries is of particular interest for future research. Linguistic features of reader production can be analyzed using natural language processing methods and potentially compared to eye-tracking data to explore the relationships between attention and language production, perhaps providing insight into the processes of developing mental models. Additionally, qualitative examination of visual data from eye-tracking was only briefly utilized in this study, but there is room to explore further the visual data from heat maps and scan paths as they allow us to witness real-time strategy use. Future studies can examine the appearance of reading strategies in visual data.

Last, motivated by the importance of intrinsic motivation in the summary score, it is important to understand how individual readers' motivation may affect their eye-movement behavior during reading. Reader perceptions, as gathered by stimulated recall, interview, or

survey, may provide further cues to aspects of readers which impact the way they approach texts. Further studies should include self-reported data from participants regarding perceptions of topic familiarity, task ease, and test authenticity which can be compared to the actual real-time reading behavior of readers.

REFERENCES

Afflerbach, P. (2016). Reading Assessment. *Reading Teacher*, *69*(4), 413–419.

Ahmed, Y., Francis, D. J., York, M., Fletcher, J. M., Barnes, M., & Kulesz, P. (2016). Validation of the direct and inferential mediation (DIME) model of reading comprehension in grades 7 through 12. *Contemporary Educational Psychology*, *44–45*, 68–82. https://doi.org/10.1016/j.cedpsych.2016.02.002

Alderson, C. J., Alderson, J. C., & Urquhart, A. H. (1984). *Reading in a Foreign Language*. Longman.

Alderson, J. C. (2000). *Assessing reading* (Atlanta Library North 4 LB1050.46 .A43 2000). Cambridge, UK ; New York, NY, USA : Cambridge University Press, 2000.

Alptekin, C., & Erçetin, G. (2010). The role of L1 and L2 working memory in literal and inferential comprehension in L2 reading. *Journal of Research in Reading*, *33*(2), 206–219. https://doi.org/10.1111/j.1467-9817.2009.01412.x

Anderson, N. J. (1999). *Exploring second language reading: Issues and strategies* (Atlanta Library North 4 PE1128.A2 A53 1999). Boston : Heinle & Heinle, ©1999.

Andrade, M. S. (2009). The Effects of English Language Proficiency on Adjustment to University Life. *International Multilingual Research Journal*, *3*(1), 16–34. https://doi.org/10.1080/19313150802668249

Artelt, C., Schiefele, U., & Schneider, W. (2001). Predictors of reading literacy. *European Journal of Psychology of Education*, *16*(3), 363–383. https://doi.org/10.1007/BF03173188

Ashby, J., Rayner, K., & Clifton, C. (2005). Eye Movements of Highly Skilled and Average Readers: Differential Effects of Frequency and Predictability. *The Quarterly Journal*

of Experimental Psychology Section A, *58*(6), 1065–1086. https://doi.org/10.1080/02724980443000476

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing; London*, *17*(1), 1–42. http://dx.doi.org.ezproxy.gsu.edu/10.1191/026553200675041464

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*(4), 453–476. https://doi.org/10.1191/0265532202lt240oa

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. OUP Oxford.

Barkaoui, K. (2014). Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL-iBT writing tasks. *Language Testing*, *31*(2), 241-259.

Barth, A., Barnes, M., Francis, D., Vaughn, S., & York, M. (2015). Inferential processing among adequate and struggling adolescent comprehenders and relations to reading comprehension. *Reading & Writing*, *28*(5), 587–609. https://doi.org/10.1007/s11145-014-9540-1

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing; London*, *30*(4), 441–465. http://dx.doi.org/10.1177/0265532212473244

Bax, S., & Chan, S. (2019). Using eye-tracking research to investigate language test validity and design. *System*, *83*, 64–78. https://doi.org/10.1016/j.system.2019.01.007

Beers, S. F., Quinlan, T., & Harbaugh, A. G. (2010). Adolescent students' reading during writing behaviors and relationships with text quality: An eyetracking study. *Part of a Special Issue: Reading during Writing. What Does Eyetracking Research Tell Us about the Interaction between Reading and Writing Processes during Text Production?*, *23*(7), 743–775. https://doi.org/10.1007/s11145-009-9193-7

Belcher, D. D., & Hirvela, A. (Eds.). (2001). *Linking literacies: Perspectives on L2 reading-writing connections*. Ann Arbor: University of Michigan Press.

Benzer, A., Sefer, A., Ören, Z., & Konuk, S. (2016). A Student-Focused Study: Strategy of Text Summary Writing and Assessment Rubric. *Education & Science / Egitim ve Bilim*, *41*(186), 163–183. https://doi.org/10.15390/EB.2016.4603

Bernhardt, E. B. (2011). *Understanding advanced second-language reading*. Routledge.

Berzak, Y., Katz, B., & Levy, R. (2018). Assessing Language Proficiency from Eye Movements in Reading. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1986–1996. http://www.aclweb.org/anthology/N18-1180

Bigelow, M., & Tarone, E. (2004). The Role of Literacy Level in Second Language Acquisition: Doesn't Who We Study Determine What We Know? *TESOL Quarterly*, *38*(4), 689–700. https://doi.org/10.2307/3588285

Booth, R. W., & Weger, U. W. (2013). The function of regressions in reading: Backward eye movements allow rereading. *Memory & Cognition*, *41*(1), 82–97. https://doi.org/10.3758/s13421-012-0244-y

Borsboom, D. (2005). *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge University Press.

Bos, L. T., De Koning, B. B., Wassenburg, S. I., & van der Schoot, M. (2016). Training Inference Making Skills Using a Situation Model Approach Improves Reading Comprehension. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.00116

Brantmeier, C. (2005). Nonlinguistic Variables in Advanced Second Language Reading: Learners' Self-Assessment and Enjoyment. *Foreign Language Annals; Alexandria, 38*(4), 494–504.

van den Broek, P., Bohn-Gettler, C., Kendeou, P., Carlson, S., & White, M. J. (2011). When a reader meets a text: The role of standards of coherence in reading comprehension. *Text Relevance and Learning from Text*. https://experts.umn.edu/en/publications/when-a-reader-meets-a-text-the-role-of-standards-of-coherence-in--2

van den Broek, P., Beker, K., & Oudega, M. (2015). Inference generation in text comprehension: Automatic and strategic processes in the construction of a mental representation. In *Inferences during reading* (pp. 94–121). Cambridge University Press.

Caccamise, D., Franzke, M., Eckhoff, A., Kintsch, E., & Kintsch, W. (2007). Guided practice in technology-based summary writing. In *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 375–396). Lawrence Erlbaum Associates Publishers.

Cain, K., Oakhill, J. V., Barnes, M. A., & Bryant, P. E. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Memory & Cognition, 29*(6), 850–859. https://doi.org/10.3758/BF03196414

Calvo, M. G. (2005). Relative contribution of vocabulary knowledge and working memory span to elaborative inferences in reading. *Learning and Individual Differences*, *15*, 53–65. https://doi.org/10.1016/j.lindif.2004.07.002

Carlson, S. E., van den Broek, P., McMaster, K., Rapp, D. N., Bohn-Gettler, C. M., Kendeou, P., & White, M. J. (2014). Effects of Comprehension Skill on Inference Generation during Reading. *International Journal of Disability, Development & Education*, *61*(3), 258–274. https://doi.org/10.1080/1034912X.2014.934004

Carrell, P. L., Carson, J. G., & Zhe, D. (1993). First and Second Language Reading Strategies: Evidence from Cloze. *Reading in a Foreign Language*, *10*(1), 953–965.

Carretti, B., Borella, E., Cornoldi, C., & De Beni, R. (2009). Role of Working Memory in Explaining the Performance of Individuals with Specific Reading Comprehension Difficulties: A Meta-Analysis. *Learning and Individual Differences*, *19*(2), 246–251.

Carver, R. P. (1997). Reading for One Second, One Minute, or One Year From the Perspective of Rauding Theory. *Scientific Studies of Reading*, *1*(1), 3.

Chateau, D., & Jared, D. (2000). Exposure to print and word recognition processes. *Memory & Cognition*, *28*(1), 143–153. https://doi.org/10.3758/BF03211582

Cheng, J., & Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Language Testing*, *35*(1), 3–25. https://doi.org/10.1177/0265532216676851

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*, 472–517. https://doi.org/10.1016/0010-0285(72)90019-9

Clifton, C., Ferreira, F., Henderson, J. M., Inhoff, A. W., Liversedge, S. P., Reichle, E. D., & Schotter, E. R. (2016). Eye movements in reading and information processing: Keith Rayner's 40year legacy. *Journal of Memory and Language*, *86*, 1–19. https://doi.org/10.1016/j.jml.2015.07.004

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

Collins, A. M., & Quillian, M. R. (1970). Facilitating retrieval from semantic memory: The effect of repeating part of an inference. *Acta Psychologica*, *33*, 304–314. https://doi.org/10.1016/0001-6918(70)90142-3

Conklin, K., Pellicer-Sanchez, A., & Carroll, G. (2018). *Eye-tracking: A guide for applied linguistics research*. Cambridge University Press.

Cook, A. E., & Wei, W. (2019). What Can Eye Movements Tell Us about Higher Level Comprehension? *Vision*, *3*(3). https://doi.org/10.3390/vision3030045

Cowan, J. R. (1976). Reading, Perceptual Strategies and Contrastive Analysis1. *Language Learning*, *26*(1), 95–109. https://doi.org/10.1111/j.1467-1770.1976.tb00262.x

Cromley, J. G., & Azevedo, R. (2007). Testing and Refining the Direct and Inferential Mediation Model of Reading Comprehension. *Journal of Educational Psychology*, *99*(2), 311–325.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing Text Readability Using Cognitively Based Indices. *TESOL Quarterly*, *42*(3), 475–493. https://doi.org/10.1002/j.1545-7249.2008.tb00142.x

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*(4), 1227–1237. https://doi.org/10.3758/s13428-015-0651-7

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting Text Comprehension, Processing, and Familiarity in Adult Readers: New Approaches to Readability Formulas. *Discourse Processes*, *54*(5–6), 340–359. https://doi.org/10.1080/0163853X.2017.1296264

Csikszentmihalyi, M. (1990). Literacy and Intrinsic Motivation. *Daedalus*, *119*(2), 115–140. JSTOR.

Cummins, J. (1979). Linguistic Interdependence and the Educational Development of Bilingual Children. *Review of Educational Research*, *49*(2), 222–251. https://doi.org/10.3102/00346543049002222

Cziko, G. A. (1978). Differences in First- and Second-Language Reading: The Use of Syntactic, Semantic and Discourse Constraints. *The Canadian Modern Language Review*, *34*(3), 473–489. https://doi.org/10.3138/cmlr.34.3.473

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*(4), 422–433. https://doi.org/10.3758/BF03214546

Davies, M. (2008). *BYU corpora: Billions of words of data: Free online access*. https://corpus.byu.edu/corpora.asp

Daza, C., & Suzuki, M. (2004). A Review of the Reading Section of the TOEIC. *TESL Canada Journal*, 16–24. https://doi.org/10.18806/tesl.v22i1.163

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension* (Atlanta Library North 4 P302 .D472 1983). New York : Academic Press, 1983.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 Reading Framework: A Working Paper*. Educational Testing Service.

Erçetin, G., & Alptekin, C. (2013). The explicit/implicit knowledge distinction and working memory: Implications for second-language reading comprehension. *Applied Psycholinguistics; New York*, *34*(4), 727–753. http://dx.doi.org/10.1017/S0142716411000932

Evans, N. W., Anderson, N. J., & Eggington, W. (2015). *ESL readers and writers in higher education: Understanding challenges, providing support*. http://public.eblib.com/choice/publicfullrecord.aspx?p=3569432

Everling, S., Gilchrist, I. D., & Liversedge, S. P. (2011). *The Oxford Handbook of Eye Movements*. OUP Oxford. http://ezproxy.gsu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=467510&site=eds-live&scope=site

Feller, D. P., Kopatich, R. D., Lech, I., & Higgs, K. (2020). Exploring Reading Strategy Use in Native and L2 Readers. *Discourse Processes*, *57*(7), 590–608. https://doi.org/10.1080/0163853X.2020.1735282

Field, J. (2018). *The cognitive validity of tests of listening and speaking designed for young learners*. Cambridge University Press. https://uobrep.openrepository.com/handle/10547/623025

Fransson, A. (1984). Cramming or understanding? Effects of intrinsic and extrinsic motivation on approach to learning and test performance. *Reading in a Foreign Language*, *4*(3), 30–54.

Gauvin, H. S., & Hulstijn, J. H. (2010). Exploring a New Technique for Comparing Bilinguals' L1 and L2 Reading Speed. *Reading in a Foreign Language*, *22*(1), 84–103.

Genesee, F., & Upshur, J. A. (1996). *Classroom-Based Evaluation in Second Language Education*. Cambridge University Press.

Godfroid, A. (2019). *Eye Tracking in Second Language Acquisition and Bilingualism: A Research Synthesis and Methodological Guide*. Routledge.

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

Grabe, W. (2010). Fluency in reading—Thirty-five years later. *Reading in a Foreign Language*, *22*(1), 71–83.

Grabe, W. P., & Stoller, F. L. (2013). *Teaching and Researching: Reading*. Routledge.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*(3), 371–395. https://doi.org/10.1037/0033-295X.101.3.371

Green, A. (2013). *Exploring Language Assessment and Testing: Language in Action*. Routledge.

Greene, B. (2001). Testing reading comprehension of theoretical discourse with cloze. *Journal of Research in Reading*, *24*, 82–98. https://doi.org/10.1111/1467-9817.00134

Guthrie, J. T., Taboada, A., & Coddington, C. S. (2007). Engagement practices for strategy learning in concept-oriented reading instruction. In *Reading Comprehension Strategies: Theories, Interventions, and Technologies* (pp. 241–266). Taylor & Francis.

Hartshorn, K. J., Evans, N. W., Egbert, J., & Johnson, A. (2017). Discipline-specific reading expectation and challenges for ESL learners in US universities. *Reading in a Foreign Language*, *29*(1), 25.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Weijer, J. van de. (2011). *Eye Tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Horiba, Y. (1996). Comprehension processes in l2 reading: Language Competence, Textual Coherence, and Inferences. *Studies in Second Language Acquisition*, *18*(4), 433–473. JSTOR.

Horiba, Y., Broek, P. W. van den, & Fletcher, C. R. (1993). Second Language Readers' Memory for Narrative Texts: Evidence for Structure-Preserving Top-Down Processing. *Language Learning*, *43*(3), 345–372. https://doi.org/10.1111/j.1467-1770.1993.tb00618.x

Hyönä, J., Lorch Jr., R. F., & Kaakinen, J. K. (2002). Individual differences in reading to summarize expository text: Evidence from eye fixation patterns. *Journal of Educational Psychology*, *94*(1), 44–55. https://doi.org/10.1037/0022-0663.94.1.44

Hyönä, J., Lorch, R., & Rinck, M. (2003). Eye Movement Measures to Study Global Text Processing. In R. Godijn, J. Theeuwes, J. Hyona, R. Radach, H. Deubel (Eds.) *The mind's eye: Cognitive and applied aspects of eye movement research.* Elsevier. https://doi.org/10.1016/B978-044451020-4/50018-9

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

Irmer, M. (2011). *Bridging Inferences: Constraining and Resolving Underspecification in Discourse Interpretation*. De Gruyter. https://doi.org/10.1515/9783110262018

Ji, N. (2011). Can a Summary Task Be Valid Writing Assessment for Less-Proficient EFL Students? *Modern Enlgish Education*, *12*(3), 46–64.

Jian, Y. (2017). Eye-movement patterns and reader characteristics of students with good and poor performance when reading scientific text with diagrams. *Reading and Writing; Dordrecht, 30*(7), 1447–1472. http://dx.doi.org.ezproxy.gsu.edu/10.1007/s11145-017-9732-6

Joh, J., & Plakans, L. (2017). Working memory in L2 reading comprehension: The influence of prior knowledge. *System, 70,* 107-120. https://doi.org/10.1016/j.system.2017.07.007

Jordan, R. R. (1997). *English for Academic Purposes: A Guide and Resource Book for Teachers*. Cambridge University Press.

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

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238. https://doi.org/10.1037/0096-3445.111.2.228

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

Just, M. A. & Carpenter, P. A. (2018, April 17). *Using Eye Fixations to Study Reading Comprehension*. In D.E. Kieras & M.A. Just (Eds.). New Methods in Reading Comprehension Research. Routledge. https://doi.org/10.4324/9780429505379-8

Kaakinen, J. K., & Hyönä, J. (2005). Perspective Effects on Expository Text Comprehension: Evidence From Think-Aloud Protocols, Eyetracking, and Recall. *Discourse Processes*, *40*(3), 239–257. https://doi.org/10.1207/s15326950dp4003_4

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading* (Atlanta Library North 4 PE1128.A2 K418 2009). Cambridge, UK ; New York : Cambridge University Press, 2009.

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge University Press.

Kintsch, W., & Yarbrough, J. C. (1982). Role of rhetorical structure in text comprehension. *Journal of Educational Psychology*, *74*(6), 828–834. https://doi.org/10.1037/0022-0663.74.6.828

Klauer, K. J., & Phye, G. D. (2008). Inductive Reasoning: A Training Approach. *Review of Educational Research*, *78*(1), 85–123. https://doi.org/10.3102/0034654307313402

Kleijn, S. (2018). *Clozing in on readability. How linguistic features affect and predict text comprehension and on-line processing.* LOT, Netherlands Graduate School.

Klichowicz, A., Scholz, A., Strehlau, S., & Krems, J. F. (2016). Differentiating between Encoding and Processing during Sequential Diagnostic Reasoning: An Eye tracking study. Presented at *Cognitive Science Society*.

Knoeferle, P., Urbach, T. P., & Kutas, M. (2011). Comprehending how visual context influences incremental sentence processing: Insights from ERPs and picture-sentence verification. *Psychophysiology*, *48*(4), 495–506. https://doi.org/10.1111/j.1469-8986.2010.01080.x

Koda, K. (1988). Cognitive Process in Second Language Reading: Transfer of L1 Reading Skills and Strategies. *Second Language Research*, *4*(2), 133–156.

Koda, K. (1990). The Use of L1 Reading Strategies in L2 Reading: Effects of L1 Orthographic Structures on L2 Phonological Recoding Strategies. *Studies in Second Language Acquisition*, *12*(4), 393–410.

Krieber, M., Bartl-Pokorny, K. D., Pokorny, F. B., Einspieler, C., Langmann, A., Körner, C., Falck-Ytter, T., & Marschik, P. B. (2016). The Relation between Reading Skills and Eye Movement Patterns in Adolescent Readers: Evidence from a Regular Orthography. *PLOS ONE*, *11*(1). https://doi.org/10.1371/journal.pone.0145934

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior research methods*, *50*(3), 1030-1046.

Lake, J. B. (2014). The Role of Individual Differences in L1 and L2 Processing of Bridging and Predictive Inferences [Thesis, Georgetown University]. In *Georgetown University-Graduate School of Arts & Sciences*. https://repository.library.georgetown.edu/handle/10822/712461

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing - LANG TEST*, *16*, 36–55. https://doi.org/10.1191/026553299672614616

Lee, J. (2011). A Comparison of Constructed Response Formats as Measures of EFL Reading Comprehension. *English Teaching*, *66*(2), 149–167.

Linacre, J. M. (2002). *What do infit and outfit, mean-square and standardized mean? Rasch Measurement Transactions, 16(2), 878*.

Linacre, J. M. (2020) Facets computer program for many-facet Rasch measurement, version 3.83.3. Beaverton, Oregon: Winsteps.com

Lipka, O., & Siegel, L. (2012). The development of reading comprehension skills in children learning English as a second language. *Reading & Writing*, *25*(8), 1873–1898. https://doi.org/10.1007/s11145-011-9309-8

LiveChat. (2016). *Typing Speed Test—Check your typing skills!*. LiveChat. https://www.livechat.com/typing-speed-test/

Loudon, C., & Macias-Muñoz, A. (2018). Item Statistics Derived from Three-Option Versions of Multiple-Choice Questions Are Usually as Robust as Four-or Five-Option Versions: Implications for Exam Design. *Advances in Physiology Education*, *42*(4), 565–575.

Macleod, C. M., Hunt, E. B., & Mathews, N. N. (1978). Individual differences in the verification of sentence—Picture relationships. *Journal of Verbal Learning and Verbal Behavior*, *17*(5), 493–507. https://doi.org/10.1016/S0022-5371(78)90293-1

Manor, B. R., & Gordon, E. (2003). Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks. *Journal of Neuroscience Methods*, *128*(1), 85–93. https://doi.org/10.1016/S0165-0270(03)00151-1

Markham, P. L. (1985). The Rational Deletion Cloze and Global Comprehension in German. *Language Learning*, *35*(3), 423–430. https://doi.org/10.1111/j.1467-1770.1985.tb01085.x

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, *99*(3), 440–466.

McNamara, D. S. (2004). SERT: Self-Explanation Reading Training. *Discourse Processes*, *38*(1), 1–30. https://doi.org/10.1207/s15326950dp3801_1

McNamara, D. S., & Magliano, J. (2009). Toward a Comprehensive Model of Comprehension. In B. Ross (Ed.). *Psychology of Learning and Motivation* (Vol. 51, pp. 297–384). Academic Press. https://doi.org/10.1016/S0079-7421(09)51009-2

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, Justice & Language assessment*. Oxford University Press.

Mokhtari, K., & Reichard, C. (2004). Investigating the strategic reading processes of first and second language readers in two different cultural contexts. *System*, *32*(3), 379–394. https://doi.org/10.1016/j.system.2004.04.005

Moss, J., Schunn, C. D., Schneider, W., McNamara, D. S., & VanLehn, K. (2011). The neural correlates of strategic reading comprehension: Cognitive control and discourse comprehension. *NeuroImage*, *58*(2), 675–686. https://doi.org/10.1016/j.neuroimage.2011.06.034

Nagy, W., Berninger, V. W., & Abbott, R. D. (2006). Contributions of morphology beyond phonology to literacy outcomes of upper elementary and middle-school students. *Journal of Educational Psychology*, *98*(1), 134–147. https://doi.org/10.1037/0022-0663.98.1.134

Nahatame, S. (2014). Strategic Processing and Predictive Inference Generation in L2 Reading. *Reading in a Foreign Language; Honolulu*, *26*(2), 54–77.

Nassaji, H. (2002). Schema Theory and Knowledge-Based Processes in Second Language Reading Comprehension: A Need for Alternative Perspectives. *Language Learning*, *52*(2), 439–481. https://doi.org/10.1111/0023-8333.00189

Noordman, L., Vonk, W., & Kempff, H. (1992). Causal inferences during the reading of expository texts. *Journal of Memory and Language*, *31*(5), 573–590. https://doi.org/10.1016/0749-596X(92)90029-W

O'Brien de Ramirez, K. (2008). *Silent, oral, L1, L2, French and English reading through eye movements and miscues* [Ph.D., The University of Arizona]. https://search.proquest.com/socialsciences/docview/304684574/abstract/BB052C32FDEB40ECPQ/57

O'Dell, F., Read, J., McCarthy, M., & Read. (2000). *Assessing Vocabulary*. Cambridge University Press.

O'Reilly, T., Feng, D. G., Sabatini, D. J., Wang, D. Z., & Gorin, D. J. (2018a). How do people read the passages during a reading comprehension test? The effect of reading purpose on text processing behavior. *Educational Assessment*, *23*(4), 277–295. https://doi.org/10.1080/10627197.2018.1513787

O'Reilly, T., Feng, D. G., Sabatini, D. J., Wang, D. Z., & Gorin, D. J. (2018b). How do people read the passages during a reading comprehension test? The effect of reading purpose on text processing behavior. *Educational Assessment*, *23*(4), 277–295. https://doi.org/10.1080/10627197.2018.1513787

Orquin, J. L., & Holmqvist, K. (2018). Threats to the validity of eye-movement research in psychology. *Behavior Research Methods*, *50*(4), 1645–1656. https://doi.org/10.3758/s13428-017-0998-z

Perfetti, C. (2007). Reading Ability: Lexical Quality to Comprehension. *Scientific Studies of Reading*, *11*, 357–383. https://doi.org/10.1080/10888430701530730

Perfetti, C. A. (1997). Sentences, individual differences, and multiple texts: Three issues in text comprehension. *Discourse Processes*, *23*(3), 337–355. https://doi.org/10.1080/01638539709544996

Pike, M. M., Barnes, M. A., & Barron, R. W. (2010). The role of illustrations in children's inferential comprehension. *Journal of Experimental Child Psychology*, *105*(3), 243–255. https://doi.org/10.1016/j.jecp.2009.10.006

Potts, G. R., Keenan, J. M., & Golding, J. M. (1988). Assessing the occurrence of elaborative inferences: Lexical decision versus naming. *Journal of Memory and Language, 27*(4), 399–415. https://doi.org/10.1016/0749-596X(88)90064-2

Prichard, C., & Atkins, A. (2016). Evaluating L2 Readers' Previewing Strategies Using Eye Tracking. *The Reading Matrix, 16*(2), 110.

Prichard, C., & Atkins, A. (2019). Selective attention of L2 learners in task-based reading online. *Reading in a Foreign Language, 31*(2), 269–290.

Raatz, U., & Klein-Braley, C. (1981). *The C-Test—A Modification of the Cloze Procedure*. In T. Culhane, C. Klein-Braley, and D. K. Stevenson, (eds.), *Practice and problems in language testing, University of Essex Department of Language and Linguistics Occasional Papers No. 26*. Colchester: University of Essex.

Ramírez, J. D. (2000). Bilingualism and Literacy: Problem or Opportunity? A Synthesis of Reading Research on Bilingual Students. *Proceedings of A Research Symposium on High Standards in Reading for Students From Diverse Language Groups: Research, Practice & Policy*, 33.

Ratcliff, R., & McKoon, G. (1978). Priming in item recognition: Evidence for the propositional structure of sentences. *Journal of Verbal Learning and Verbal Behavior*, *17*(4), 403–417. https://doi.org/10.1016/S0022-5371(78)90238-4

Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin*, *85*(3), 618–660. https://doi.org/10.1037/0033-2909.85.3.618

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). *Eye movements as reflections of comprehension processes in reading. Scientific*. 241–255.

Rayner, K., McConkie, G. W., & Zola, D. (1980). Integrating information across eye movements. *Cognitive Psychology*, *12*, 206–226. https://doi.org/10.1016/0010-0285(80)90009-2

Ridgway, T. (1994). Reading Theory and Foreign Language Reading Comprehension. *Reading in a Foreign Language*, *10*(2), 55–83.

Rodriguez, M. C. (2005). Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13.

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

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology*, *25*(1), 54–67. https://doi.org/10.1006/ceps.1999.1020

Schaffner, E., & Schiefele, U. (2013). The prediction of reading comprehension by cognitive and motivational factors: Does text accessibility during comprehension testing make a difference? *Learning and Individual Differences*, *26* (Supplement C), 42–54. https://doi.org/10.1016/j.lindif.2013.04.003

Segers, E., & Verhoeven, L. (2016). How logical reasoning mediates the relation between lexical quality and reading comprehension. *Reading and Writing*, *29*(4), 577–590. https://doi.org/10.1007/s11145-015-9613-9

Seidlhofer, B. (1990). Summary Judgments: Perspectives on Reading and Writing. *Reading in a Foreign Language*, *6*(2), 413–424.

Shimizu, H. (2009). The Effects of Causal Relatedness on EFL Learners' Reading Comprehension and Inference Generation. *ARELE: Annual Review of English Language Education in Japan, 20*, 31–40. https://doi.org/10.20581/arele.20.0_31

Singer, M., Halldorson, M., Lear, J. C., & Andrusiak, P. (1992). Validation of causal bridging inferences in discourse understanding. *Journal of Memory and Language, 31*(4), 507–524. https://doi.org/10.1016/0749-596X(92)90026-T

Snow, C. (2002). *Reading for Understanding: Toward an R&D Program in Reading Comprehension*. Rand Corporation.

Spivey, N. N. (1990). Transforming Texts: Constructive Processes in Reading and Writing. *Written Communication*, *7*(2), 256–287. https://doi.org/10.1177/0741088390007002004

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly*, *16*(1), 32–71.

Stoller, F. L., Anderson, N. J., Grabe, W., & Komiyama, R. (2013). Instructional Enhancements to Improve Students' Reading Abilities. *English Teaching Forum, 51*(1), 2-11,.

Tarchi, C. (2015). Fostering reading comprehension of expository texts through the activation of readers' prior knowledge and inference-making skills. *International Journal of Educational Research, 72*, 80–88. https://doi.org/10.1016/j.ijer.2015.04.013

Taylor, J. N., & Perfetti, C. A. (2016). Eye movements reveal readers' lexical quality and reading experience. *Reading and Writing; Dordrecht*, *29*(6), 1069–1103. http://dx.doi.org/10.1007/s11145-015-9616-6

Taylor, L. (2013). *Testing Reading through Summary: Investigating summary completion tasks for assessing reading comprehension ability*. Cambridge University Press.

Ulijn, J. M., & Kempen, G. a. M. (1976). *The role of the first language in second hand language reading reading comprehension: Some experimental evidence*. https://research.tue.nl/en/publications/the-role-of-the-first-language-in-second-hand-language-reading-re

Urquhart, A. H., & Alderson, J. Charles. (1984). *Reading in a foreign language*. Longman.

Urquhart, A. H., & Weir, C. J. (2014). *Reading in a Second Language: Process, Product and Practice*. Routledge.

Verhoeven, L., Leeuwe, J. van, & Vermeer, A. (2011). Vocabulary Growth and Reading Development across the Elementary School Years. *Scientific Studies of Reading*, *15*(1), 8–25. https://doi.org/10.1080/10888438.2011.536125

Wallot, S. (2011). *The role of reading fluency, text difficulty and prior knowledge in complex reading tasks.* University of Cincinnati.

Wang, J., Li, L., Li, S., Xie, F., Chang, M., Paterson, K. B., White, S. J., & McGowan, V. A. (2018). Adult Age Differences in Eye Movements During Reading: The Evidence From Chinese. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *73*(4), 584–593. https://doi.org/10.1093/geronb/gbw036

Wang, Z., Sabatini, J., O'Reilly, T., & Feng, G. (2017). How Individual Differences Interact With Task Demands in Text Processing. *Scientific Studies of Reading*, *21*(2), 165–178. https://doi.org/10.1080/10888438.2016.1276184

Wigfield, A., & Guthrie, J. T. (1997). Relations of children's motivation for reading to the amount and breadth or their reading. *Journal of Educational Psychology*, *89*(3), 420–432. https://doi.org/10.1037/0022-0663.89.3.420

Williams, R. S., Ari, O., & Santamaria, C. N. (2011). Measuring college students' reading comprehension ability using cloze tests. *Journal of Research in Reading*, *34*(2), 215–231. https://doi.org/10.1111/j.1467-9817.2009.01422.x

Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch measurement transactions*, *8*(3), 370.

Yamashita, J. (2008). Extensive reading and development of different aspects of L2 proficiency. *System*, *36*(4), 661–672. https://doi.org/10.1016/j.system.2008.04.003

Yeari, M., Elentok, S., & Schiff, R. (2017). Online and offline inferential and textual processing of poor comprehenders: Evidence from a probing method. *Journal of Experimental Child Psychology*, *155*(Supplement C), 12–31. https://doi.org/10.1016/j.jecp.2016.10.011

Zwaan, R. A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin & Review*, *23*(4), 1028–1034. https://doi.org/10.3758/s13423-015-0864-x

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185. https://doi.org/10.1037/0033-2909.123.2.162

APPENDICES

Appendix A Demographic survey

Please enter your age.

_____

Please enter your gender.

_____

What do you consider to be your first language(s)?

_____

In what country did you attend high school?

_____

What was the language of instruction at your high school?

_____

At what age did you begin to learn English?

_____

How long have you lived in a country where English is the most spoken language?

_____

How many classes have you taken for the purpose of learning English since you began learning?

_____

On what occasions do you use English (in class, at home, reading online, etc.)?

_____

Appendix B English morpho-syntactic knowledge and vocabulary size measure.

In each sentence, finish the word with the blank (___). Every word with a blank is missing half of the letters in the word.

1. More than half the houses are occu_____ by renters.

(Answer: Occupied. Target: vocabulary and passive syntax)

2. The camera fl_____ captured the pose before the bride could breathe or blink.

(Answer: Flash. Target: vocabulary)

3. It's so biz_____, I can't even really comprehend it.

(Answer: bizarre. Target: vocabulary)

4. To show their loyalty, workers are putting in extra hours to bo_____ production.

(Answer: boost. Target: vocabulary)

5. Just now we've conf_____ reports of a tornado on the ground near the city.

(Answer: confirmed. Target: vocabulary and inflectional morphology for aspect)

6. There was simply no room to st_____ them.

(Answer: store. Target: vocabulary)

7. Realizing his mistake, he bac_____ slowly up the block checking each doorway.

(Answer: backed. Target: tense)

8. If we conti_____ say things that are untrue, we will not be believed.

(Answer: continually. Target: vocabulary and derivational morphology)

9. The woman wears a black scarf wra_____ around her neck and head.

(Answer: wrapped. Target: inflectional morphology)

10. He said the company will coop_____ fully with authorities.

(Answer: cooperate. Target: vocabulary)

11. His father's family can tr_____ its history in the region back to the 1700s.

(Answer: trace. Target: vocabulary)

12. That lawsuit was dism_____ earlier this year, but the family is appealing the case.

(Answer: dismissed. Target: vocabulary and passive syntax)

13. She couldn't hear a thing over the poun_____ of her heart.

(Answer: pounding. Target: vocabulary and inflectional morphology for aspect)

14. She wor_____ about the consequences, especially now after the serious injury.

(Answer: worried. Target: vocabulary and tense)

15. The judge's decision could de_____ the trial for months.

(Answer: delay. Target: vocabulary)

16. He telephoned for an appointment with a psychi_____ whom his doctor had recommended to him.

(Answer: psychiatrist. Target: vocabulary and derivational morphology)

17. Through the window, a band of sunlight stre_____ across the room.

(Answer: streamed. Target: vocabulary and tense)

18. You would have to look back a very long time to find a historical prec_____.

(Answer: precedent. Target: vocabulary)

Appendix C Reading Motivation Survey.

Instructions: Please indicate your level of agreement with the following statements, zero (0) indicating no agreement and four (4) indicating very strong agreement.

1. I often read as quickly as I can to get only useful information.
2. I believe good reading skills are key to success.
3. I often prefer reading to other activities during my free time.
4. I believe reading skills have little value in one's profession.
5. I like to talk about things I learned through reading.
6. I have a difficult time staying interested in reading material.
7. I am most committed to reading something when it is a mandatory.
8. I see reading only as a tool for accomplishing school- and work-related tasks.
9. I enjoy following up discussions on new topics by reading more about them.
10. I look forward to reading challenging books and texts.

Extrinsic motivation questions: 1, 2, 4 (reverse scored), 5, 7

Intrinsic motivation questions: 3, 6 (reverse scored), 8 (reverse scored), 9, 10

Appendix D Reading comprehension test forms

Multiple-choice forms:

"Biotechnology"

Biotechnology is defined as the making of useful products by using living systems and organisms. These products are part of medicine, agriculture, food production, to name a few. Biotechnology permeates all parts of our lives and asks us to make important ethical decisions at times. DNA Technology is one of the areas of biotechnology that is centered around the use of DNA.

The Human Genome Project, or HGP, was an international effort to determine all the base pairs of the human genome. It is the world's largest collaborative biological project to date, beginning in 1990 and having completed in 2003. It also aimed to map all the genes discovered onto their respective chromosomes. Among other applications and benefits, knowing the human genome's code allows us to discover the source of diseases and design effective treatments. The HGP's public database is used by scientists exploring other DNA Technologies.

Scientists utilize the genetic "fingerprints" or profiles of humans to analyze DNA evidence. DNA profiles are the sets of unique letters that make up a person's genome. To create someone's fingerprint, their genome is broken into pieces that target parts of DNA that vary greatly among humans, since humans are 99.9% identical otherwise. The pieces are separated on a gel in bands. The pattern of bands is unique to individuals, and related persons share common bands. Forensic scientists use DNA profiles in criminal investigation. DNA profiles can also be used to determine paternity.

Genetic engineering, in a general sense, is any modification of an organism's DNA by using biotechnology. It may involve knocking out genes, inserting genes or even targeting specific genes with an intended mutation within an organism. There are a variety of ways in which genetic engineering may be used. One of its uses is for molecular cloning. This involves using an organism, such as bacteria, as a protein factory. This is how we are able to manufacture enzymes for use in detergents, as well as produce large amounts of insulin or human growth hormone for human medical uses.

Multiple-choice questions:

1. The main topic of this passage is: (Main Idea)
a. The relationship between biotechnology and fingerprints.
b. The applications of biotechnology related to DNA.
c. The medical uses of genetic engineering.

2. What was the goal of the Human Genome Project (HGP)? (Detail)
a. To map the human genetic code for scientific applications
b. To develop better biotechnology from genomes
c. To complete the world’s largest collaborative biological project

3. Why are only some parts of human DNA useful for researchers? (Bridging Inference)
a. Many DNA patterns do not appear in the HGP database.
b. It is possible to modify an organism’s DNA by using biotechnology.
c. Very little human DNA is unique to one person.

4. Which of the following is true about molecular cloning? (detail)
a. It can be done in a protein factory.
b. It results in medical products for human use.
c. It is the first step in cloning larger organisms.

5. Based on the article, which of the following is a potential use of DNA technology? (Elaborative Inference)
a. Removing genes which are known to be harmful.
b. Identifying someone based on fingerprints.
c. Molecular cloning of human beings.

Answer key
1: b
2: a
3: c
4: b
5: a

"Microscope"

The compound microscope uses a series of lenses in order to magnify an image so that the subtle characteristics of that object are more clearly seen. Historically, the development of the compound microscope has been attributed to several people. Perhaps the most famous and accepted history of the modern compound microscope is that of Galileo Galilei who is said to have developed a compound microscope with adjustable focus in 1609.

The compound microscope works by gathering light, redirecting it through a condenser lens and into the path of the specimen. The condenser lens focuses or condenses the light onto the specimen and is needed for higher magnification because it increases the illumination of the light and the resolution. The image of the specimen is then directed to the back portion of the microscope, called the focal plane, by the objective lens. The image from the focal plane is then received by the ocular lens and the image is redirected to the eye. Once the image reaches the eye, it is actually viewed in reverse of its orientation on the slide; essentially the image is upside down and backwards from the orientation on the stage. A compound microscope can generally magnify a specimen in a range of about 40X to 400X but could be magnified up to 1000X in some compound microscopes.

While the compound microscope is very useful, it is also limited by its resolution. Resolution is the shortest distance between two separate points in a microscope's field of view that can still be distinguished as distinct entities. It directly relates to the clarity of the image when viewed. If the image lacks resolution, it will appear "fuzzy" and individual components or characteristics of the image may be obscured.

Still, the compound microscope is an essential part of crime labs for very small or dense pieces of evidence. Compound microscopes are most useful when high magnification is needed but are limited by the size of the object to be viewed. The item must be small enough to fit on slides on the stage while still fitting under the objective lenses.

Multiple-choice questions:

1. Which statement best describes the main idea of the text? (Main idea)
a) How the modern compound microscope functions
b) How to increase resolution of the compound microscope
c) What components constitute the compound microscope

2. Where is an image first directed to in the compound microscope? (Detail)
a) the focal plane
b) the condenser lens
c) the objective lens

3. What does the condenser lens do? (Detail)
a) Reflects and refracts light onto the specimen
b) Focuses the light onto the specimen
c) Receives an image from the focal plane

4. When is a microscope’s quality best? (Bridging Inference)
a) When the magnification is large
b) When the resolution is high.
c) When the object being magnified is small.

5. A compound microscope would be most useful for: (Elaborative Inference)
a) performing surgery
b) examining bacteria
c) making computer software

Answer key
1: a
2: a
3: b
4: b
5: b

"Water"

The importance of water to life, and therefore biology, cannot be understated. It covers over seventy percent of the Earth and is the most abundant compound in living things. All living things on Earth depend upon water to survive. Water is required for many essential reactions within cells, such as cell respiration and photosynthesis.

Water is a simple but unique molecule that is tasteless, odorless, and transparent. Its chemical formula is H_2O. It has hydrogen atoms that are covalently bonded to an oxygen atom. What makes water unique, and so important for life, are the interesting characteristics, or properties, that water displays as a result of its structure.

Water is a neutral molecule, meaning that it has the same number of protons as electrons. Even though water is neutral, its electrons are unequally distributed among the oxygen and hydrogens that make it up. The oxygen atom, with its eight positively charged protons, has a strong pull on the negatively charged electrons; this makes the probability of finding those electrons near the oxygen greater than finding them near the hydrogen atoms. Water is therefore a polar molecule.

Water's polarity also makes it a very good solvent. This is biologically helpful because it means that water can transport or hold onto

dissolved substances for organisms (salt, food). Because water is polar itself, when it comes into contact with other polar or ionic substances, it is able to fit in between the atoms that make up that substance, dissolving it. In other words, these substances can mix. Salt or rubbing alcohol will dissolve in water and are therefore called hydrophilic, or "water loving." Water cannot dissolve non- polar substances, such as oil, or fats, and will often show a separation from them acting as if it is "squeezing" them together. This is called the hydrophobic effect ("water fearing"). This effect is very important in the formation of cell membranes.

Multiple-choice questions:

1. What is the main purpose of this text? (main idea)
a. To explain why water is essential for living things
b. To explain how water dissolves substances in living things
c. To explain the abundance of water on Earth

2. Where are electrons located in a water molecule? (detail)
a. Mostly on hydrogen atoms
b. Mostly on oxygen atoms
c. The same on the oxygen and hydrogens that make it up.

3. How does water’s polarity make it useful? (detail)
a. It allows water to dissolve substances such as salt and food.
b. It allows organisms to take in more oxygen.
c. It allows oxygen atoms and hydrogen atoms to mix.

4. Which of the following is TRUE about water? (bridging inference)
a. Water is unique, and therefore one of the most abundant molecules in living things.
b. Water is neutral, which means it has very few noticeable physical characteristics.
c. Water is a solvent, and thus it is good for transporting substances throughout the body.

5. According to the passage, salt is what kind of substance? (bridging inference)
a. ionic
b. hydrophobic
c. non-polar

Answer key
1: a
2: b
3: a
4: c
5: a

“Hunger”

You are biologically motivated to eat to survive, but also by psychological, social, and cultural factors, which makes hunger motivation an interesting study. One factor in hunger motivation is certainly biology. When your stomach is empty, it alerts you to a need for food. But psychological research shows that your empty stomach isn't the only factor that leads to hunger. When participants' stomachs were filled with inflated balloons, they did feel hunger eventually, although their stomachs continued to experience fullness because of the balloons.

Hunger, then, does not come from the stomach, but from the brain. One specific structure in the brain, the hypothalamus, regulates feelings of hunger and fullness. The lateral hypothalamus creates feelings of hunger and the ventromedial hypothalamus creates feelings of fullness. When functioning correctly, the hypothalamus senses appetite hormones in the blood stream and creates a balance, keeping the body at a comfortable weight, or set point.

If body weight increases, the hypothalamus decreases hunger and increases metabolic rate to get back to the "normal weight" that the body has been at for a period of time. If body weight decreases, the hypothalamus increases hunger and decreases metabolic rate to get back to the normal weight. This explains why it's so difficult for people to diet and lose weight; if the body has been at a high weight for a period of time, the body

starts to think that's the "normal" weight and seeks to remain at that weight.

This biological explanation does not fully explain hunger motivation though. While some people are motivated by internal cues (hunger hormones or a growling stomach), others are motivated by external cues such as stress or the smell or sight of something that appeals to them. Intrinsic and extrinsic motivation apply not only to hunger motivation, but to our motivation in other areas of life as well. Culture also affects our motivation to eat specific foods. The foods that we grow up eating become familiar and desirable to us, and new foods are often viewed with disgust.

Multiple-choice questions:

1. What is the main idea of this passage? (Main idea)
a. Culture is an important factor in determining our hunger.
b. Hunger is motivated by biological, psychological, and cultural factors.
c. Biological motivation for survival causes us to get hungry.

2. Which of the following statements are true about the hypothalamus? (detail)
a. The hypothalamus makes hormones in the mind and in the blood.
b. The hypothalamus contains multiple parts to create the feeling of hunger.
c. The hypothalamus can control how our bodies lose and gain weight.

3. Why is it difficult for people to lost weight? (detail)
a. If someone has a higher weight for a long time, the hypothalamus thinks a higher weight is normal.
b. The hypothalamus increases a person’s hunger and decreases their metabolism.
c. Higher weights lead to increased hunger motivation in the brain.

4. Why might someone feel hungry even if they have eaten recently? (Elaborative Inference)
a. The hypothalamus triggers hunger because body weight has increased.
b. They did not eat the right food to create a feeling of fullness.
c. They notice food which reminds them of a good memory.

5. The balloon experiment is mentioned to show which of the following? (Bridging Inference)
a. People can experience fullness even without eating food.
b. People feel hunger based on biological signals from the stomach.
c. People can still experience hunger even if they are full.

Answer key
1: b
2: c
3: a
4: c
5: c

"Choices"

All choices mean that one alternative is selected over another. Selecting among alternatives involves three ideas central to economics: scarcity, choice, and opportunity cost. Using the economy's scarce resources to produce one thing requires giving up another. Producing better education, for example, may require cutting back on other services, such as health care. A decision to preserve a wilderness area requires giving up other uses of the land. Every society must decide what it will produce with its scarce resources.

There are not many free goods. Outer space, for example, was a free good when the only use we made of it was to gaze at it. But now, our use of space has reached the point where one use can be an alternative to another. Conflicts have already arisen over the allocation of orbital slots for communications satellites. Thus, even parts of outer space are scarce. Space will surely become scarcer as we find new ways to use it. Scarcity characterizes virtually everything.

Opportunity cost is what you missed out on getting when you chose to do something else. The cost can be in dollars, time, or anything. If you had to

choose between going to college or getting a job, the opportunity cost of choosing college would be the money you have to spend on tuition plus the money you would have earned if you had a job. If you chose getting a job, the opportunity cost would be the diploma you would have earned in college, all the things you would have learned, and all the friends you would have made. We use opportunity cost to determine which choice is the better one.

A trade-off is a situation that involves losing one quality or aspect of something in return for gaining another quality or aspect. More colloquially, if one thing increases, some other thing must decrease. In economics, a trade-off is commonly expressed in terms of the opportunity cost of one potential choice, which is the loss of the best available alternative. The concept of a trade-off is often used to describe situations in everyday life.

Multiple-choice questions:

1. Which of the following could be a good title for this passage? (Main idea)

A. Society's use of scarce resources
B. Scarcity, Choice, and Opportunity Cost
C. The art of making choices

2. Why does the author include the information about Outer space? (Detail)

a. To explain that our use of space is an example of using a free good.
b. To support the idea that scarcity is a common property of things.
c. To hypothesize that space will become scarcer if we use it in more ways.

3. Which of the following is true about opportunity costs? (Detail)

A. Opportunity cost is calculated by combining the cost of both choices.
B. We must pay opportunity costs before we can make profitable choices.
C. Thinking about opportunity cost means comparing the potential loss behind each choice.

4. What is the relationship between opportunity cost and a trade-off? (Bridging Inference)
A. The two terms refer to the same phenomenon but are used in different occasions.

B. Making a choice in a trade-off situation involves paying the opportunity cost of the choice.

C. In economics, if one thing increases in a trade-off, some other thing must decrease.

5. Which of the following would be part of the opportunity cost of choosing to live by yourself versus living with a roommate? (Elaborative Inference)
A. Avoiding arguing with a roommate about chores and bills.
B. The money that would be saved by sharing the rental cost.
C. Quality of life gained by choosing to not live with a roommate.

Answer key
1: b
2: b
3: c
4: b
5: b

"Attitudes"

You may be surprised to find that your actions can affect your attitudes. This tendency for actions to affect attitudes can be explained through cognitive dissonance theory, which says that people experience dissonance, or uncomfortable tension, when their actions and attitudes don't match, and because they can't undo their actions, they have to change their attitude to relieve the tension. If you think it's wrong to talk badly about people, but one day you gossip about a friend, you might feel uncomfortable because your action doesn't match your attitude.

You might decide it's okay to talk badly about friends under certain circumstances. And according to Cognitive Dissonance Theory, now it will be easier for you to talk badly about friends in the future because your attitude has changed. Sometimes the role you play, as student or employee or boyfriend or scientist, can also affect your attitudes. When you start your new role, you're careful to follow the social expectations, although you might feel like you're acting.

One famous social psychology study performed by Stanford Professor Philip Zimbardo showed how playing a role can affect attitudes. In his experiment, Zimbardo randomly assigned college students to act as either prisoners or prison guards. The guards were given mirrored sunglasses, uniforms, and clubs, and they were asked to take charge of the prisoners. The prisoners were forced to wear nightgown-type outfits and kept in prison-cell type rooms in the basement floor of a college building.

Within a couple of days of playing these roles, the students assigned to be prison guards started to create cruel and degrading practices (forcing prisoners to wear bags over their heads to travel down the hall to the bathroom or forcing prisoners to sleep without blankets). Students assigned to play prisoners experienced emotional breakdowns or passively resigned themselves to the bad treatment. Zimbardo was forced to end the experiment after only six days because the students' behaviors were so out-of-control. Acting out the role

led to a change in students' attitudes and led to more intense

actions.

Multiple-choice comprehension questions:

1. What is most likely the main idea of the text? (Main idea)

a) People feel uncomfortable when their actions don’t match.
b) Social expectations can change a person’s personality.
c) People’s attitudes can change to justify actions.

2. Which of the following is true according to Cognitive Dissonance Theory? (Detail)

a) Acting out roles allows people to justify behavior.
b) Social roles allow people to talk badly about their friends.
c) Our actions do not match our psychological expectations.

3. What was the purpose of Professor Zimbardo’s study? (Detail)

a) To study psychological problems in the prison system
b) To show how attitudes are affected by social roles
c) To prove the relevance of social psychology

4. Which of the following lessons about psychology were learned from the Zimbardo Study? (Bridging Inference)

a) Behaviors can be adjusted depending on social roles.
b) Using students as subject may ruin experiments.
c) Prisons are cruel and degrading places.

5. How can knowledge of cognitive dissonance theory help people? (Elaborative Inference)

a) Mental discomfort can be reduced through passive acceptance.
b) Behavior change might become harder when problems are uncomfortable.
c) People experiencing dissonance make mental adjustments to reduce discomfort.

Answer Key
1: c
2: a
3: b
4: a
5: c

Cloze forms:

"Biotechnology"

Biotechnology is defined as the making of useful products by using living systems and organisms. These _____ are part of medicine, agriculture, food production, to name a few. Biotechnology permeates all parts of our lives and asks us to make important ethical _____ at times. DNA Technology is one of the areas of biotechnology that is _____ around the use of DNA.

The Human Genome Project, or HGP, was an international effort to _____ all the base pairs of the human genome. It is the world's largest collaborative biological _____ to date, beginning in 1990 and having completed in 2003. It also _____ to map all the genes discovered onto their respective chromosomes. Among other applications and benefits, _____ the human genome's code allows us to discover the source of _____ and design effective treatments. The HGP's public database is used by scientists exploring other DNA Technologies.

Scientists _____ the genetic "fingerprints" or profiles of humans to analyze DNA evidence. DNA _____ are the sets of unique letters that make up a person's genome. To create someone's fingerprint, their genome is _____ into pieces that target parts of DNA that vary greatly among humans, since humans are 99.9% identical otherwise. The pieces are separated on a gel in bands. The pattern of _____ is unique to individuals, and related persons share _____ bands. Forensic scientists use DNA profiles in criminal investigation. DNA profiles can also be used to determine paternity.

Genetic engineering, in a general sense, is any _____ of an organism's DNA by using biotechnology. It may involve _____ out genes, inserting genes or even targeting specific genes with an intended mutation within an organism. There are a variety of ways in which genetic engineering may be used. One of its uses is for molecular cloning. This involves using an organism, such as bacteria, as a protein factory. This is how we are able to manufacture enzymes for use in detergents, as well as produce large amounts of insulin or human growth hormone for human medical uses.

“Microscope”

The compound microscope uses a series of lenses in order to magnify an image so that the subtle characteristics of that object are more clearly seen. Historically, the _____ of the compound microscope has been attributed to several people. Perhaps the most famous and accepted _____ of the modern compound microscope is that of Galileo Galilei who is said to have developed a compound microscope with adjustable focus in 1609.

The compound microscope works by _____ light, redirecting it through a condenser lens and into the path of the specimen. The condenser lens _____ or condenses the light onto the specimen and is needed for higher magnification because it increases the illumination of the light and the resolution. The _____ of the specimen is then directed to the back portion of the microscope, called the focal plane, by the objective lens. The image from the focal plane is then _____ by the ocular lens and the image is redirected to the eye. Once the image reaches the eye, it is actually _____ in reverse of its orientation on the slide; essentially the image is upside down and backwards from the _____ on the stage. A compound microscope can generally magnify a specimen in a _____ of about 40X to 400X but could be magnified up to 1000X in some compound microscopes.

While the compound microscope is very _____, it is also limited by its resolution. Resolution is the shortest _____ between two separate points in a microscope's field of view that can still be distinguished as distinct entities. It _____ relates to the clarity of the image when viewed. If the image lacks resolution, it will _____ "fuzzy" and individual components or characteristics of the image may be obscured.

Still, the compound microscope is an _____ part of crime labs for very small or dense pieces of evidence. Compound microscopes are most useful when high magnification is needed _____ are limited by the size of the object to be viewed. The item must be small enough to fit on slides on the stage while still fitting under the objective lenses.

"Water"

The importance of water to life, and therefore biology, cannot be understated. It covers over 70% of the Earth and is the most abundant compound in _____ things. All living things on Earth _____ upon water to survive. Water is required for many essential reactions within cells, such as cell respiration and photosynthesis.

Water is a simple _____ unique molecule that is tasteless, odorless, and transparent. Its chemical formula is H_2O. It has hydrogen atoms that are covalently bonded to an oxygen _____. What makes

water unique, and so important for life, are the interesting characteristics, or properties, that water _____ as a result of its structure.

Water is a neutral molecule, meaning that it has the same _____ of protons as electrons. Even though water is neutral, its electrons are unequally distributed among the oxygen and _____ that make it up. The oxygen atom, with its eight positively charged protons, has a strong pull on the _____ charged electrons; this makes the probability of finding those electrons near the oxygen greater than finding them _____ the hydrogen atoms. Water is therefore a polar molecule.

Water's polarity also _____ it a very good solvent. This is biologically helpful _____ it means that water can transport or hold onto dissolved substances for organisms (salt, food). Because water is polar itself, when it comes into _____ with other polar or ionic substances, it is able to fit in between the atoms that make up that substance, dissolving it. In other words, these substances can mix. Salt or rubbing alcohol will _____ in water and are therefore called hydrophilic, or "water loving." Water cannot dissolve non-polar _____, such as oil, or fats, and will often show a separation from them acting as if it is "squeezing" them together. This is called the hydrophobic effect ("water fearing"). This _____ is very important in the formation of cell membranes.

"Hunger"

You are biologically motivated to eat to survive, but also by psychological, social, and cultural factors, which makes hunger motivation an interesting study. One factor in hunger motivation is certainly biology. When your stomach is _____, it alerts you to a need for food. But psychological research _____ that your empty stomach isn't the only factor that leads to hunger. When participants' stomachs were _____ with inflated balloons, they did feel hunger eventually, _____ their stomachs continued to experience fullness because of the balloons.

Hunger, then, does not come from the stomach, but from the _____. One specific structure in the brain, the hypothalamus, _____ feelings of hunger and fullness. The lateral hypothalamus creates feelings of hunger and the ventromedial hypothalamus _____ feelings of fullness. When functioning correctly, the hypothalamus senses appetite hormones in the blood stream and creates a balance, _____ the body at a comfortable weight, or set point.

If body weight increases, the hypothalamus decreases hunger and _____ metabolic rate to get back to the "normal weight" that the body has been at for a period of time. If body weight decreases, the hypothalamus increases _____ and decreases metabolic rate to

get back to the normal weight. This explains why it's so difficult for people to diet and _____ weight; if the body has been at a high weight for a period of time, the body starts to think that's the "normal" weight and seeks to _____ at that weight.

This biological explanation does not fully _____ hunger motivation though. While some people are motivated by internal cues (hunger hormones or a growling stomach), others are motivated by _____ cues such as stress or the smell or sight of something that appeals to them. Intrinsic and extrinsic motivation apply not only to hunger motivation, _____ to our motivation in other areas of life as well. Culture also affects our motivation to eat specific foods. The foods that we grow up eating become familiar and desirable to us, and new foods are often viewed with disgust.

"Choices"

All choices mean that one alternative is selected over another. Selecting among alternatives involves three ideas central to economics: scarcity, choice, and opportunity cost. Using the economy's scarce resources to produce one thing _____ giving up another. Producing better education, for _____, may require cutting back on other services, such as health care. A decision to preserve a wilderness area requires giving up other uses of the land. Every society must decide what it will _____ with its scarce resources.

There are not many free goods. Outer space, for example, was a _____ good when the only use we made of it was to gaze at it. _____ now, our use of space has reached the point where one use can be an _____ to another. Conflicts have already arisen over the allocation of orbital slots for communications satellites. Thus, even parts of outer space are scarce. Space will surely _____ scarcer as we find new ways to use it. Scarcity characterizes virtually everything.

Opportunity cost is what you missed out on getting when you _____ to do something else. The _____ can be in dollars, time, or anything. If you had to choose _____ going to college or getting a job, the opportunity cost of choosing college would be the _____ you have to spend on tuition plus the money you would have earned if you had a job. _____ you chose getting a job, the opportunity cost would be the diploma you would have earned in college, all the things you would have learned, and all the friends you would have made. We use opportunity cost to determine which _____ is the better one.

A trade-off is a situation that involves _____ one quality or aspect of something in return for gaining another quality or aspect. More colloquially, if one thing increases, some other thing must decrease. In economics, a trade-off is commonly expressed in terms of the _____ cost of one potential choice, which is the loss of

the best available alternative. The concept of a trade-off is often used to describe situations in everyday life.

"Attitudes"

You may be surprised to find that your actions can affect your attitudes. This tendency for _____ to affect attitudes can be explained through cognitive dissonance theory, which says that people experience dissonance, or uncomfortable tension, when their actions and attitudes don't match, and _____ they can't undo their actions, they have to change their attitude to relieve the tension. If you think it's wrong to talk badly about people, but one day you gossip about a friend, you might _____ uncomfortable because your action doesn't match your attitude.

You might decide it's okay to _____ badly about friends under certain circumstances. And according to Cognitive Dissonance Theory, now it will be _____ for you to talk badly about friends in the future because your attitude has changed. Sometimes the _____ you play, as student or employee or boyfriend or scientist, can also affect your attitudes. _____ you start your new role, you're careful to follow the social expectations, although you might feel like you're acting.

One famous social psychology study _____ by Stanford Professor Philip Zimbardo showed how playing a role can affect

attitudes. In his experiment, Zimbardo randomly assigned college students to _____ as either prisoners or prison guards. The guards were _____ mirrored sunglasses, uniforms, and clubs, and they were asked to take charge of the prisoners. The prisoners were forced to _____ nightgown-type outfits and kept in prison-cell type rooms in the basement floor of a college building.

Within a couple of days of playing these roles, the students assigned to be prison _____ started to create cruel and degrading practices (forcing prisoners to wear bags over their heads to travel down the hall to the bathroom or forcing prisoners to _____ without blankets). Students assigned to play _____ experienced emotional breakdowns or passively resigned themselves to the bad treatment. Zimbardo was forced to end the experiment after only six days because the students' behaviors were so _____. Acting out the role led to a change in students' attitudes and led to more intense actions.

Summary Prompt*:

Imagine you are in a class on the above subject, and you are assigned to teach a fellow student about the content of this text. Write a summary using 50 to 150 words.

*This prompt was used for every summary form. The text which accompanied it was the same as the one for multiple-choice question tasks above.

Appendix E Sentences used in sentence verification tasks

Sentence	Related-condition text	Unrelated-condition text	Veracity	Number of words (characters)
Medical technology is rarely seen as controversial.	“Biotechnology”	“Microscope”	FALSE	7 (51)
Most human fingerprints are nearly identical to each other.	“Biotechnology”	“Microscope”	FALSE	9 (59)
Proteins and enzymes are created by only large bacteria.	“Biotechnology”	“Microscope”	FALSE	9 (56)
Genes change naturally throughout an average person’s life.	“Biotechnology”	“Microscope”	FALSE	9 (60)
Some diseases are related to a person’s genetics.	“Biotechnology”	“Microscope”	TRUE	9 (56)
Related people share similar genetic information.	“Biotechnology”	“Microscope”	TRUE	9 (50)
Every living thing contains unique genetic information.	“Biotechnology”	“Microscope”	TRUE	6 (49)
Great scientific efforts involve international cooperation.	“Biotechnology”	“Microscope”	TRUE	7 (55)
Scientists only observe things that we can see normally.	“Microscope”	“Biotechnology”	FALSE	6 (59)
Light and sound waves never change direction after hitting an object.	“Microscope”	“Biotechnology”	FALSE	9 (56)
The history of scientific technology spans only a few years.	“Microscope”	“Biotechnology”	FALSE	11 (69)
A microscope tells us things about large objects far away.	“Microscope”	“Biotechnology”	FALSE	10 (60)
Scientific procedures require precise and accurate data.	“Microscope”	“Biotechnology”	TRUE	10 (58)
Visible light allows us to see the objects around us.	“Microscope”	“Biotechnology”	TRUE	8 (60)
Higher quality images are clearer than low quality images.	“Microscope”	“Biotechnology”	TRUE	7 (56)
Common pieces of technology involve complex parts inside.	“Microscope”	“Biotechnology”	TRUE	10 (53)
Every atom and molecule contain balanced protons and electrons.	“Water”	“Hunger”	FALSE	9 (58)
Water is an element containing multiple smaller molecules.	“Water”	“Hunger”	FALSE	8 (57)

A phobia is a strong love for a specific thing.	“Water”	“Hunger”	FALSE	9 (63)
Liquids like oil are similar enough to water to mix with it.	“Water”	“Hunger”	FALSE	8 (58)
Living creatures need to drink water to help with digestion.	“Water”	“Hunger”	TRUE	10 (47)
Different types of molecules each have a unique structure.	“Water”	“Hunger”	TRUE	12 (60)
All living things on Earth are made of cells.	“Water”	“Hunger”	TRUE	10 (60)
One way two substances can mix is through dissolving.	“Water”	“Hunger”	TRUE	9 (58)
Anatomy explains everything related to human life.	“Hunger”	“Water”	FALSE	9 (45)
The human brain consists of one uniform organ.	“Hunger”	“Water”	FALSE	9 (53)
It is impossible to change a persons body shape.	“Hunger”	“Water”	FALSE	7 (50)
A human’s childhood has little impact on adult behavior.	“Hunger”	“Water”	FALSE	8 (46)
People remember specific feelings when they sense specific input.	“Hunger”	“Water”	TRUE	10 (50)
People often relate being healthy to weighing less.	“Hunger”	“Water”	TRUE	10 (57)
Biologists and psychologists often study different things.	“Hunger”	“Water”	TRUE	9 (65)
Our brains control the way our bodies function.	“Hunger”	“Water”	TRUE	8 (51)
More competition for a resource makes it less scarce.	“Choices”	“Attitudes”	FALSE	7 (58)
Principles of economics only apply to big businesses.	“Choices”	“Attitudes”	FALSE	8 (47)
All involved parties benefit in a trade-off of choices.	“Choices”	“Attitudes”	FALSE	9 (53)
Governments need not worry about the use of scarce resources.	“Choices”	“Attitudes”	FALSE	8 (53)
Success in college requires a large time commitment.	“Choices”	“Attitudes”	TRUE	9 (55)
A free good is available for anyone to use as much as possible.	“Choices”	“Attitudes”	TRUE	10 (61)
There is a finite amount of money in the world.	“Choices”	“Attitudes”	TRUE	8 (52)

Countries must compete for space and resources.	“Choices”	“Attitudes”	TRUE	13 (63)
Researchers are not responsible for treating human subjects well.	“Attitudes”	“Choices”	FALSE	10 (47)
People consider it healthy to gossip about others.	“Attitudes”	“Choices”	FALSE	7 (47)
Humans often act based on pure logic over their beliefs.	“Attitudes”	“Choices”	FALSE	9 (65)
Prisoners are treated better than guards in most prisons.	“Attitudes”	“Choices”	FALSE	8 (50)
Tough experiences can cause long-term mental issues.	“Attitudes”	“Choices”	TRUE	10 (56)
Repeated actions become increasingly easy to do.	“Attitudes”	“Choices”	TRUE	9 (57)
The human mind adapts to deal with difficult situations.	“Attitudes”	“Choices”	TRUE	7 (52)
Jobs come with specific rules to be followed.	“Attitudes”	“Choices”	TRUE	7 (48)

Appendix F Summary rating guidelines given to raters

For the summary tasks, participants were shown a text to read, and then asked to write a summary of the text. The participants were instructed to write the summary as if it was directed at a fellow classmate who missed the assignment, and the summary needed to be between 50 and 150 words.

As raters, you will see the source text, the prompt for writing a summary, and the respondent's summary. The source texts are from the same pool of 6 texts as in the cloze texts, so you may run into texts you have seen when rating the cloze tests. You will have space to rate the summary and leave comments or questions.

When rating the summaries, make sure you are familiar with the source text (please read through the text fully the first time you encounter a source text). After you comprehend the source text, rate the participants' summaries on four constructs: accuracy, modeling, task completion, and language use:

- **Accuracy** – This relates to how well a summary accurately reflects the topic and propositions of the source text. Reporting information which is correct with respect to the text increases accuracy. Inclusion of propositions which are incorrect with respect to the source or show evidence of misunderstanding decrease accuracy. Accuracy could also be decreased by major omissions of ideas from large portions of the source text.
- **Modeling** – Modeling refers to how well a reader can read across an entire text and create a condensed mental model of the text. For summaries, this relates to how well a summary captures the main idea of the text, avoids irrelevant or trivial information, and generalizes across smaller details. Well-modelled summaries show a balance of brevity and detail. Use of generalizations capturing multiple points, and reliance on statements which relate to the main

idea and topic of the source increase modeling. Use of propositions related to minute, trivial, and irrelevant information, and omission of major propositions, decrease modeling.

- **Task completion** – This relates to how well the summary meets the requirements for the task. The following are the requirements:
    - summary should be structured for the intended audience. The intended audience was instructed to be the participants' classmates who may have some knowledge of the subject matter.
    - within the word limit (between 50 and 150 words)
    - written in an academic register
    - Ideas are organized in a coherent fashion. It should read like a text and not disconnected ideas.

  Brevity, coherent organization, and audience mindfulness increase task completion. Disorganization, major slighting of source information, and neglect of the audience and register decrease task completion. Summaries which are too short or too long are considered lower in task completion.
- **Language use** – This relates to how accurate the writer uses grammar and vocabulary and if the summary is not written in an overly informal register. Evidence of paraphrasing over direct copying increases language use. Grammatical accuracy, use of sophisticated and relevant vocabulary, and successful paraphrasing also increase language use. Grammatical inaccuracies, misused words, and direct copying decrease language use.

Other Considerations:

- It is important to assign ratings to each construct as independently as possible. For example, please try not to make judgments about summary Accuracy based on Language Use (although some overlap is inevitable).
- **KEEP IN MIND:** Because of constrains of the test interface, typos and punctuations were very difficult for participants to correct. Therefore, errors related strictly to spelling and punctuation are likely to be frequent even in high-quality summaries, and these errors should not be considered toward any part of the summary score.

The rubric is presented below. Each construct is to be evaluated separately on a scale from 0 to 4. Only use 0 if no evidence of the relevant construct is present in the summary.

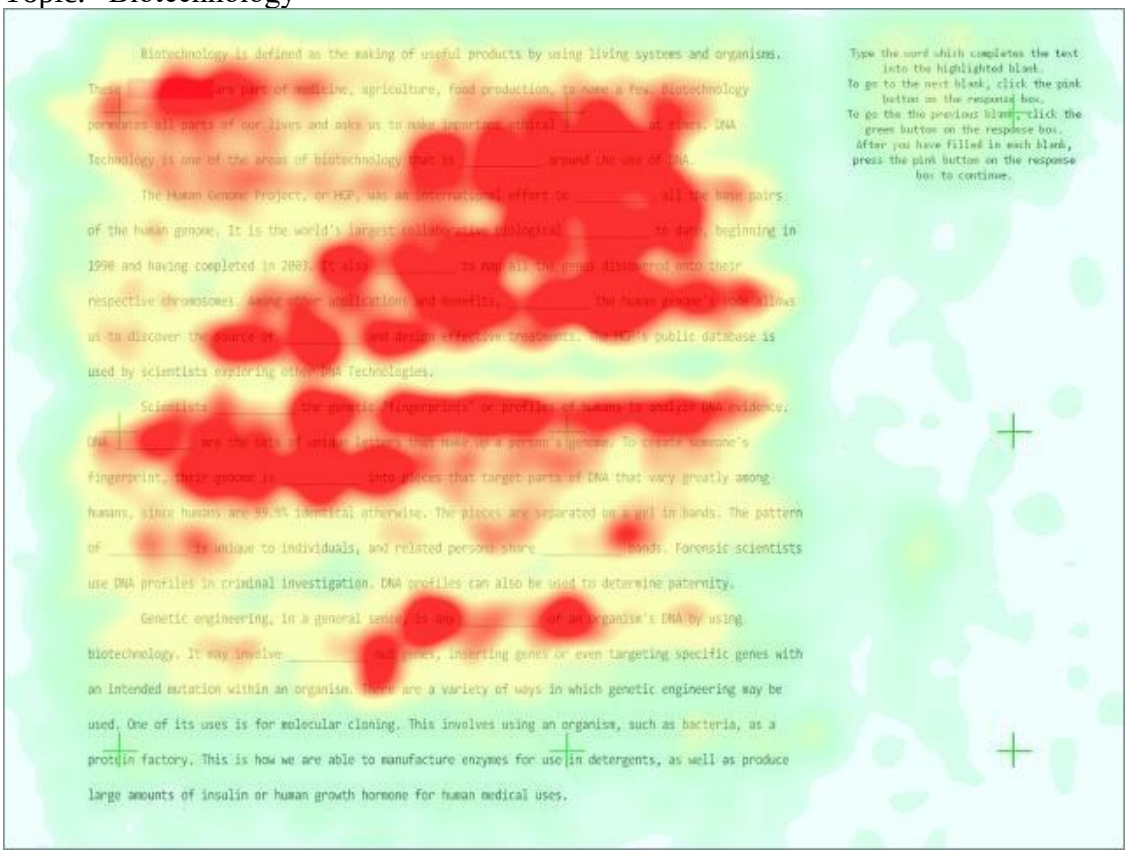
Score	Accuracy	Modeling	Task completion	Language use
4	The summary accurately reflects the topic and propositions in the source text. Misinterpretations of information are few and may be due to wording and not to misunderstanding of the source text. No main ideas are outright omitted.	The summary captures the main idea of the text, while avoiding irrelevant or trivial information. Substantial amounts of smaller details are generalized into briefer propositions.	The summary is organized in an appropriate way for the intended audience: fellow classmates. The summary communicates the ideas of the text coherently, and within the word limit.	Writer uses a wide range of lexical and syntactic structures that may go beyond the wording in the source. Ideas are appropriately reformulated and not directly lifted. Few, if any, errors.
3	Somewhat accurate account of the text. Propositions from text generally reflect source content and topic. No main ideas are outright omitted.	The summary focuses on main ideas of the text, but may focus on some trivial or irrelevant aspects of the source. Conversely, important information may be slighted.	The summary is well-organized, but may include too much or too little detail, which may affect its effectiveness for the purposes of the given task.	Writer uses a wide range of syntactic and lexical structures, with appropriate levels of paraphrasing of ideas. Syntactic and lexical errors are few.

2	Shows some understanding of text content, although there may be some distortions or omissions which affect accuracy. May have at least one major error of comprehension.	The summary gives equal focus to the text's main ideas and subordinate ideas. The author may highlight ideas which do not capture the larger points of the text.	Fair level of coherence, although content may be disorganized. The length of the summary is either under 50 words, or well over 150 words.	Some use of original wording, but there are examples of verbatim or near-copy uses of source text. Many syntactic and lexical errors may be present.
1	Little evidence of understanding the text. Propositions are mostly inaccurate with respect to the source text and do not capture the topic of the text.	The summary shows minimal evidence that the writer has created an accurate mental model of the text. The summary is not focused on main ideas and does little to combine subordinate ideas.	The summary does not fit within the length requirements of the task, has little attempt at organizing ideas, and does not relate the information from the source text in an academic fashion.	Writer shows very basic understanding of vocabulary and syntactic structures, with heavy reliance on verbatim copying of source language.
0	No evidence of comprehension. Summary is off-topic or has no relevant facts.			

Appendix G Heat maps showing aggregate intensity (number and duration) of fixations in each task-topic condition.

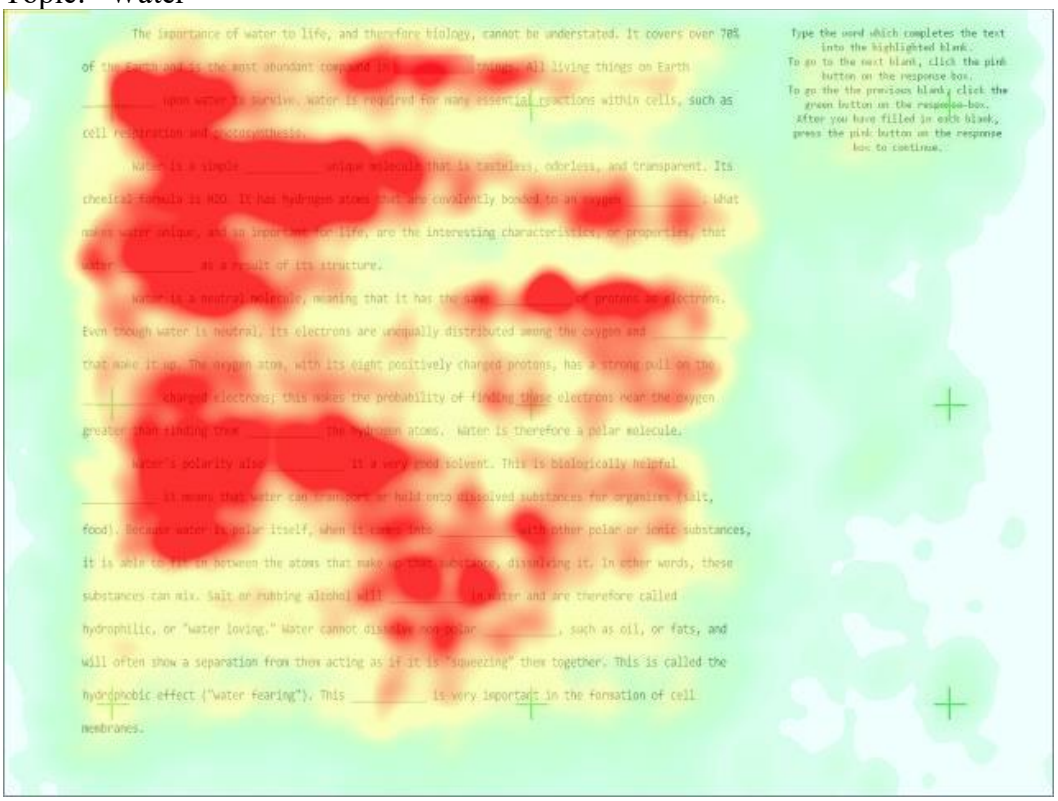
Cloze Tasks:

Topic: “Biotechnology”

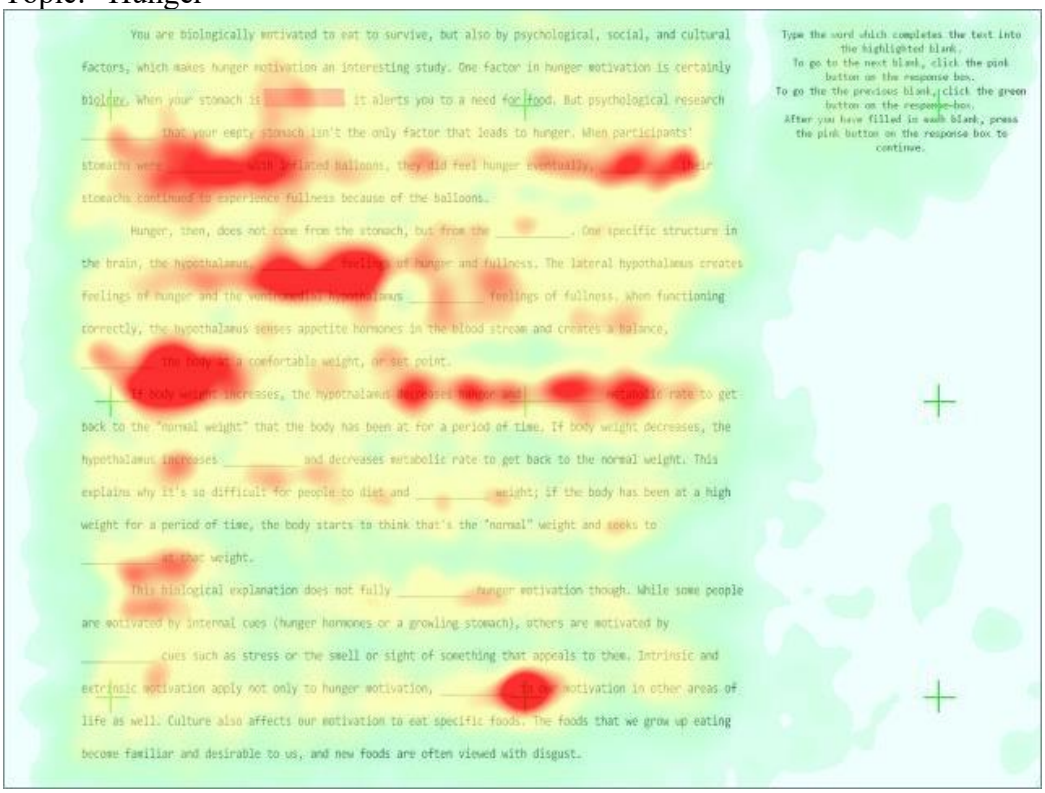


Topic: “Compound Microscope”

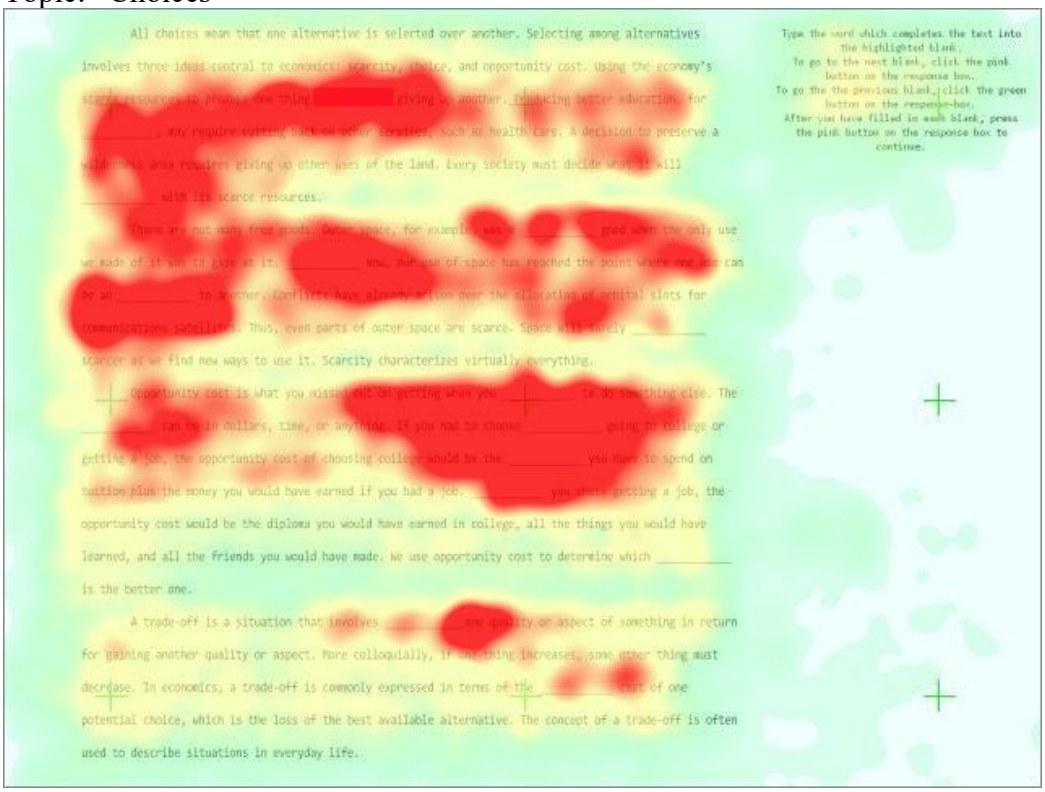
Topic: “Water”



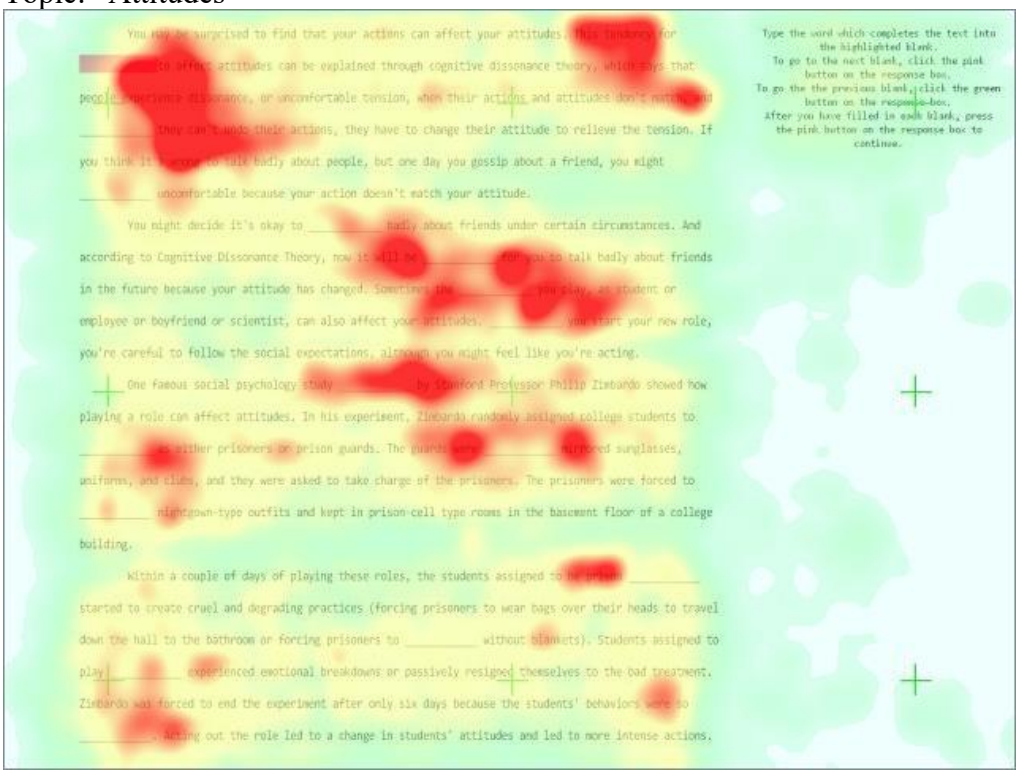
Topic: “Hunger”



Topic: “Choices”

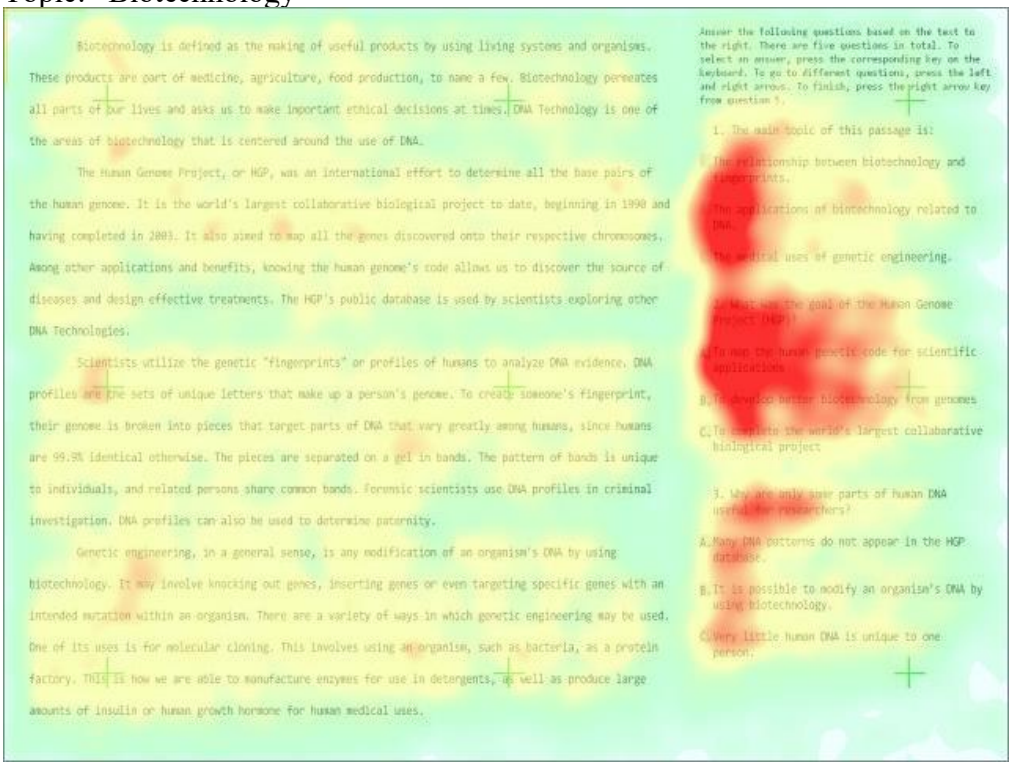


Topic: “Attitudes”

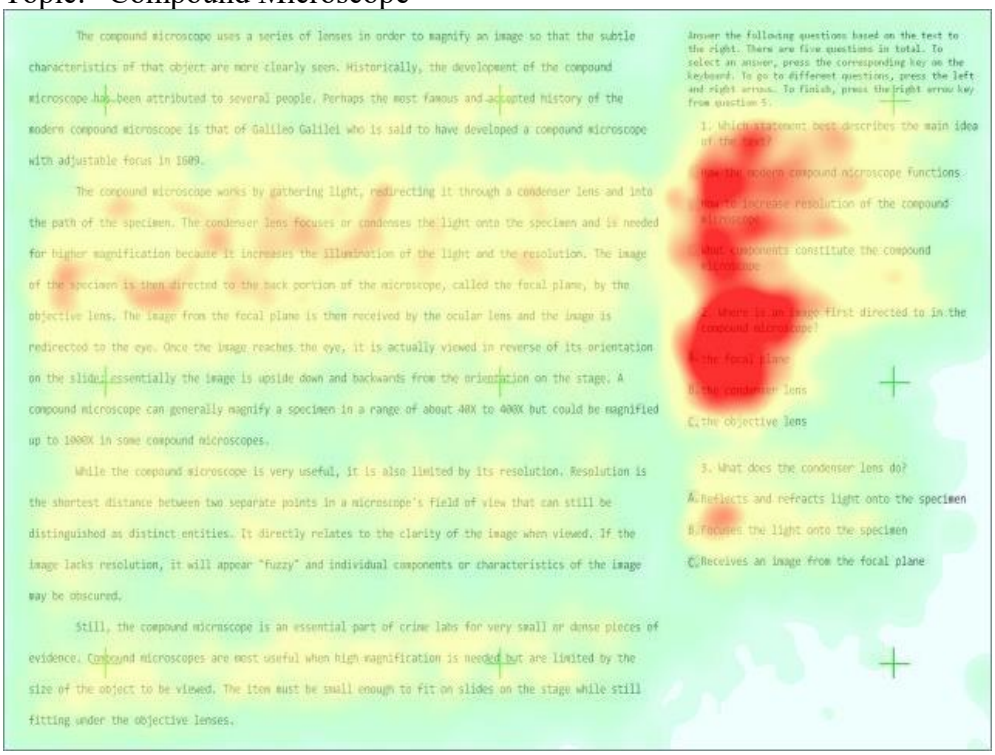


MC Tasks:

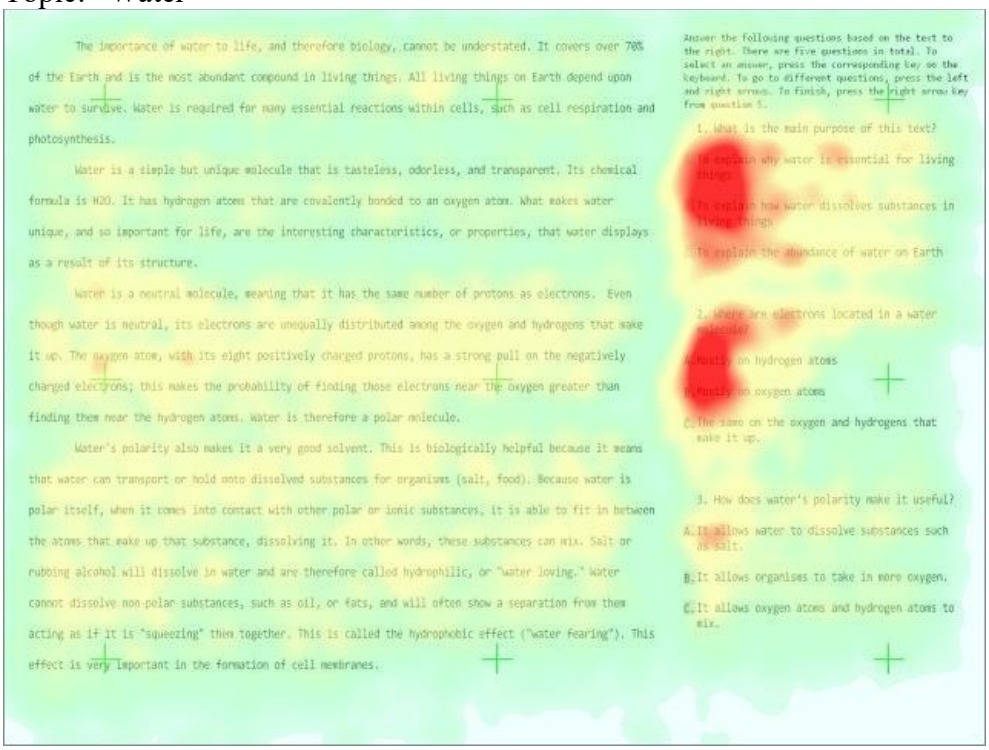
Topic: “Biotechnology”



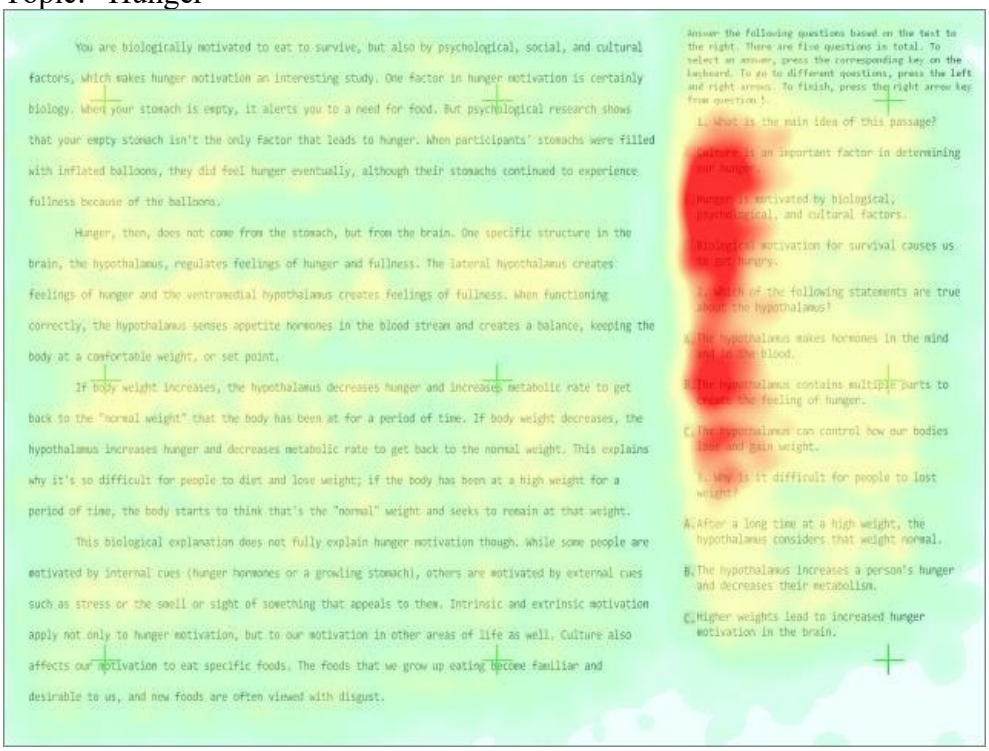
Topic: “Compound Microscope”



Topic: “Water”



Topic: “Hunger”



You are biologically motivated to eat to survive, but also by psychological, social, and cultural factors, which makes hunger motivation an interesting study. One factor in hunger motivation is certainly biology. When your stomach is empty, it alerts you to a need for food. But psychological research shows that your empty stomach isn't the only factor that leads to hunger. When participants' stomachs were filled with inflated balloons, they did feel hunger eventually, although their stomachs continued to experience fullness because of the balloons.

Hunger, then, does not come from the stomach, but from the brain. One specific structure in the brain, the hypothalamus, regulates feelings of hunger and fullness. The lateral hypothalamus creates feelings of hunger and the ventromedial hypothalamus creates feelings of fullness. When functioning correctly, the hypothalamus senses appetite hormones in the blood stream and creates a balance, keeping the body at a comfortable weight, or set point.

If body weight increases, the hypothalamus decreases hunger and increases metabolic rate to get back to the "normal weight" that the body has been at for a period of time. If body weight decreases, the hypothalamus increases hunger and decreases metabolic rate to get back to the normal weight. This explains why it's so difficult for people to diet and lose weight; if the body has been at a high weight for a period of time, the body starts to think that's the "normal" weight and seeks to remain at that weight.

This biological explanation does not fully explain hunger motivation though. While some people are motivated by internal cues (hunger hormones or a growling stomach), others are motivated by external cues such as stress or the smell or sight of something that appeals to them. Intrinsic and extrinsic motivation apply not only to hunger motivation, but to our motivation in other areas of life as well. Culture also affects our motivation to eat specific foods. The foods that we grow up eating become familiar and desirable to us, and new foods are often viewed with disgust.

Answer the following questions based on the text to the right. There are five questions in total. To select an answer, press the corresponding key on the keyboard. To go to different questions, press the left and right arrows. To finish, press the right arrow key from question 5.

1. What is the main idea of this passage?

A. Culture is an important factor in determining your hunger.

B. Hunger is motivated by biological, psychological, and cultural factors.

C. Biological motivation for survival causes us to get hungry.

2. Which of the following statements are true about the hypothalamus?

A. The hypothalamus makes hormones in the mind and in the blood.

B. The hypothalamus contains multiple parts to create the feeling of hunger.

C. The hypothalamus can control how our bodies lose and gain weight.

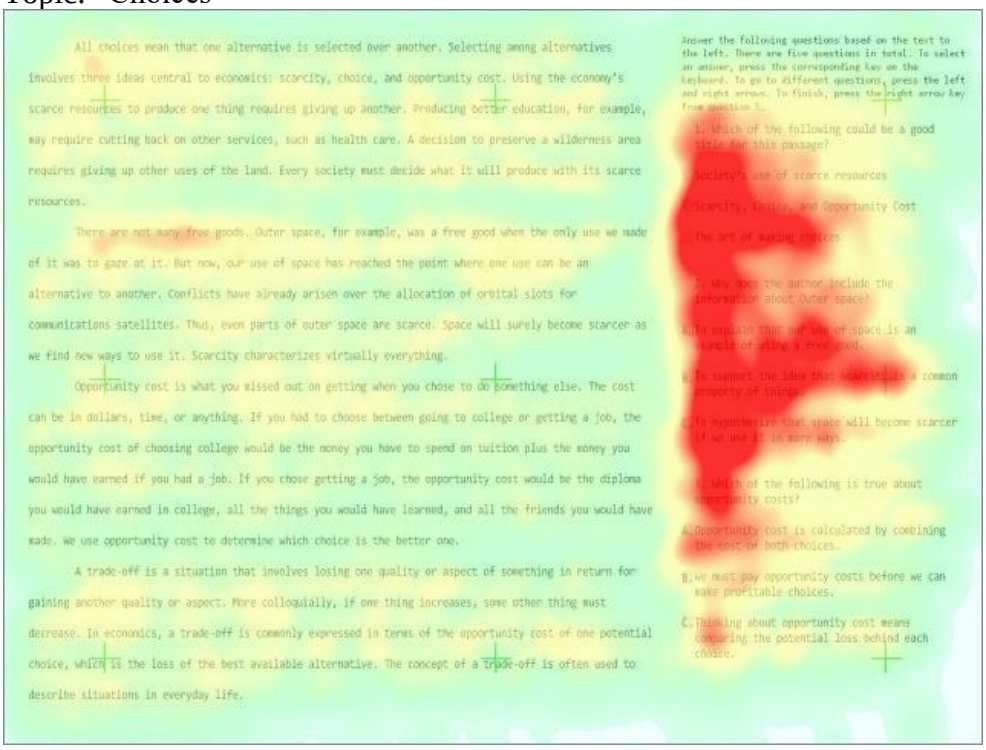
3. Why is it difficult for people to lost weight?

A. After a long time at a high weight, the hypothalamus considers that weight normal.

B. The hypothalamus increases a person's hunger and decreases their metabolism.

C. Higher weights lead to increased hunger motivation in the brain.

Topic: “Choices”



All choices mean that one alternative is selected over another. Selecting among alternatives involves three ideas central to economics: scarcity, choice, and opportunity cost. Using the economy's scarce resources to produce one thing requires giving up another. Producing better education, for example, may require cutting back on other services, such as health care. A decision to preserve a wilderness area requires giving up other uses of the land. Every society must decide what it will produce with its scarce resources.

There are not many free goods. Outer space, for example, was a free good when the only use we made of it was to gaze at it. But now, our use of space has reached the point where one use can be an alternative to another. Conflicts have already arisen over the allocation of orbital slots for communications satellites. Thus, even parts of outer space are scarce. Space will surely become scarcer as we find new ways to use it. Scarcity characterizes virtually everything.

Opportunity cost is what you missed out on getting when you chose to do something else. The cost can be in dollars, time, or anything. If you had to choose between going to college or getting a job, the opportunity cost of choosing college would be the money you have to spend on tuition plus the money you would have earned if you had a job. If you chose getting a job, the opportunity cost would be the diploma you would have earned in college, all the things you would have learned, and all the friends you would have made. We use opportunity cost to determine which choice is the better one.

A trade-off is a situation that involves losing one quality or aspect of something in return for gaining another quality or aspect. More colloquially, if one thing increases, some other thing must decrease. In economics, a trade-off is commonly expressed in terms of the opportunity cost of one potential choice, which is the loss of the best available alternative. The concept of a trade-off is often used to describe situations in everyday life.

Answer the following questions based on the text to the left. There are five questions in total. To select an answer, press the corresponding key on the keyboard. To go to different questions, press the left and right arrows. To finish, press the right arrow key from question 5.

1. Which of the following could be a good title for this passage?

A. Society's use of scarce resources

B. Scarcity, Choice, and Opportunity Cost

C. The art of making choices

2. Why does the author include the information about outer space?

A. To explain that parts of space is an example of being a free good.

B. To support the idea that scarcity is a common property of things.

C. To hypothesize that space will become scarcer if we use it in more ways.

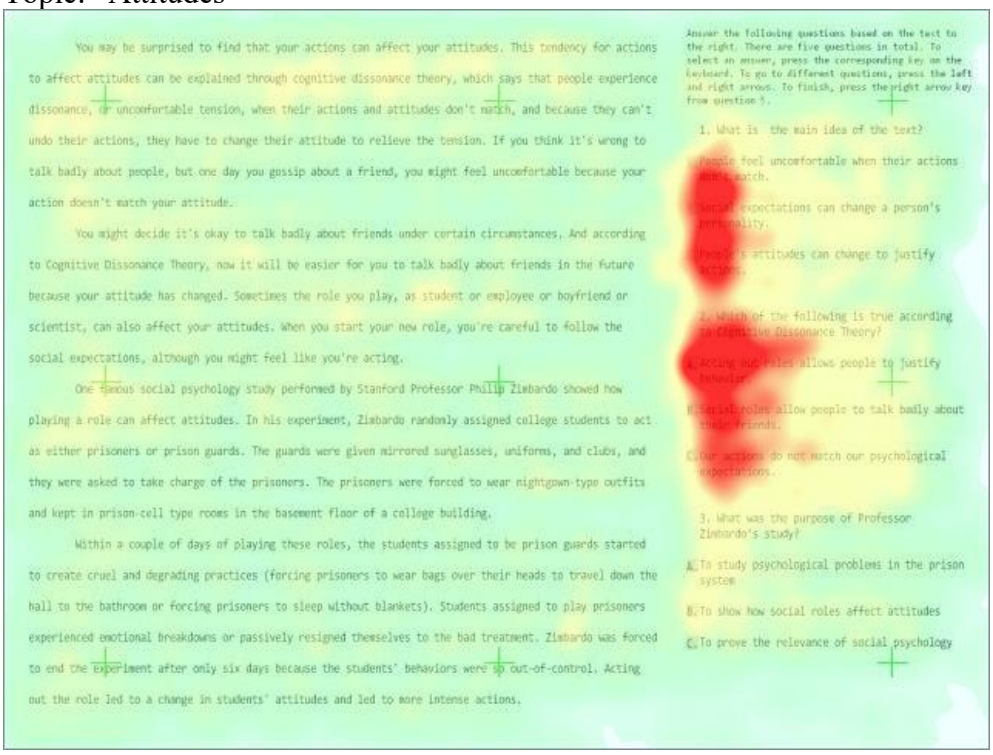
3. Which of the following is true about opportunity costs?

A. Opportunity cost is calculated by combining the cost of both choices.

B. We must pay opportunity costs before we can make profitable choices.

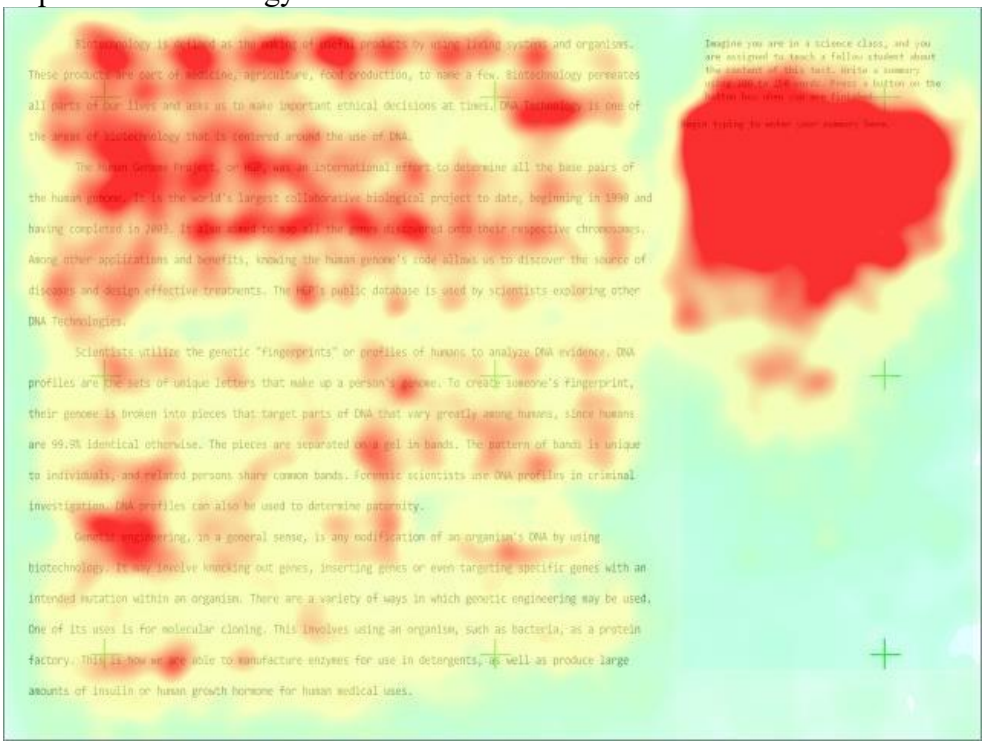
C. Thinking about opportunity cost means considering the potential loss behind each choice.

Topic: “Attitudes”

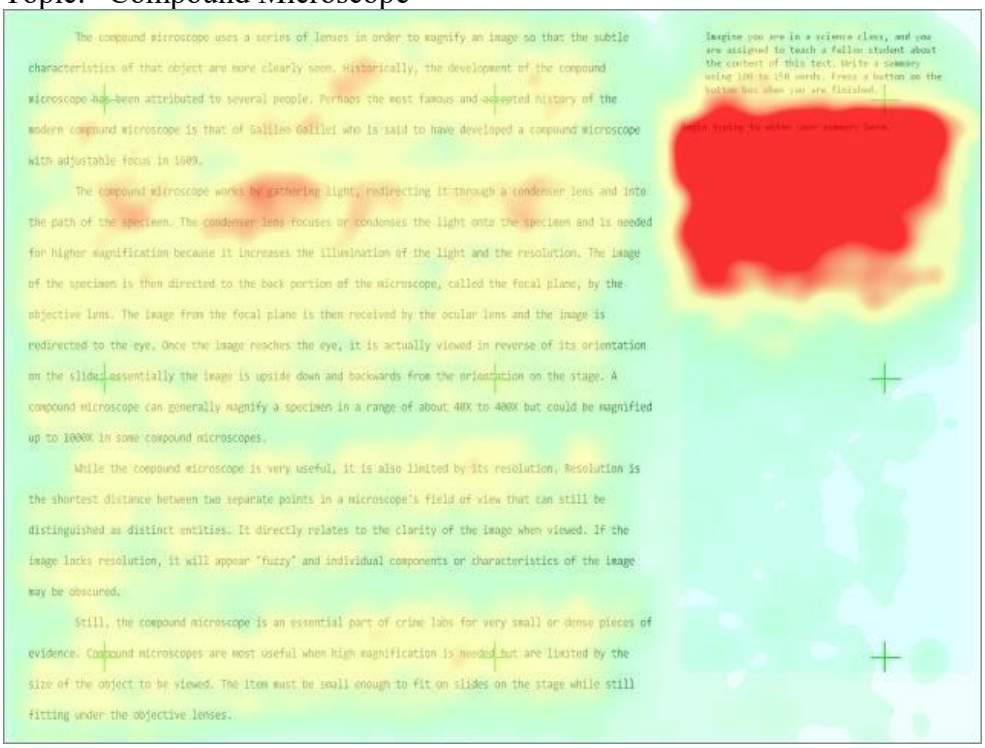


Summary Tasks:

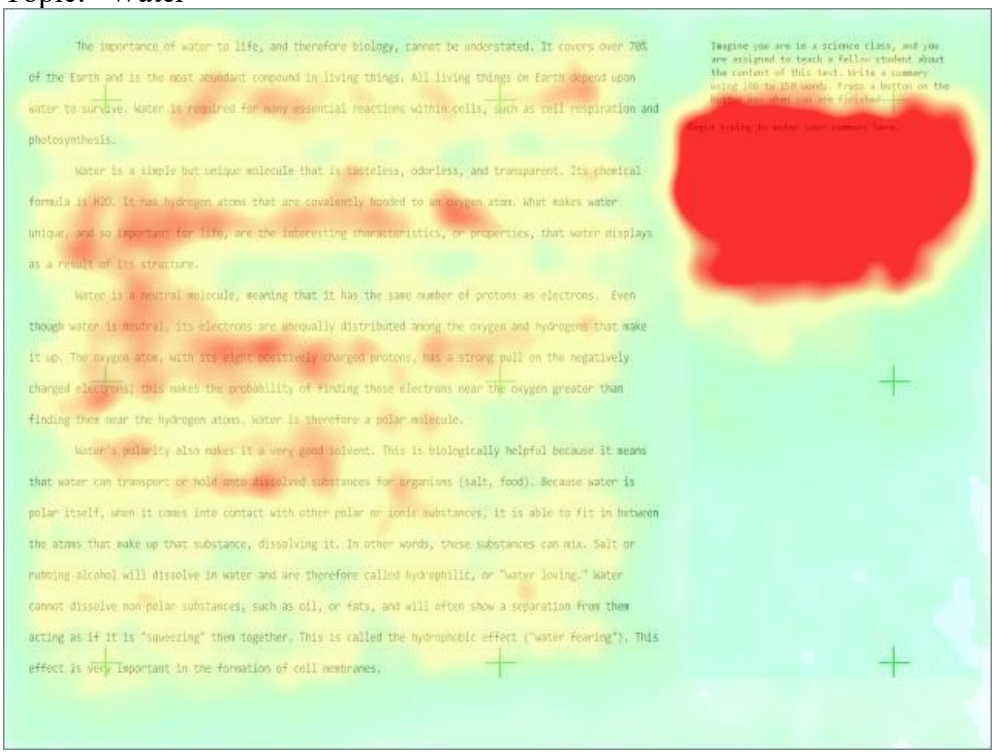
Topic: “Biotechnology”



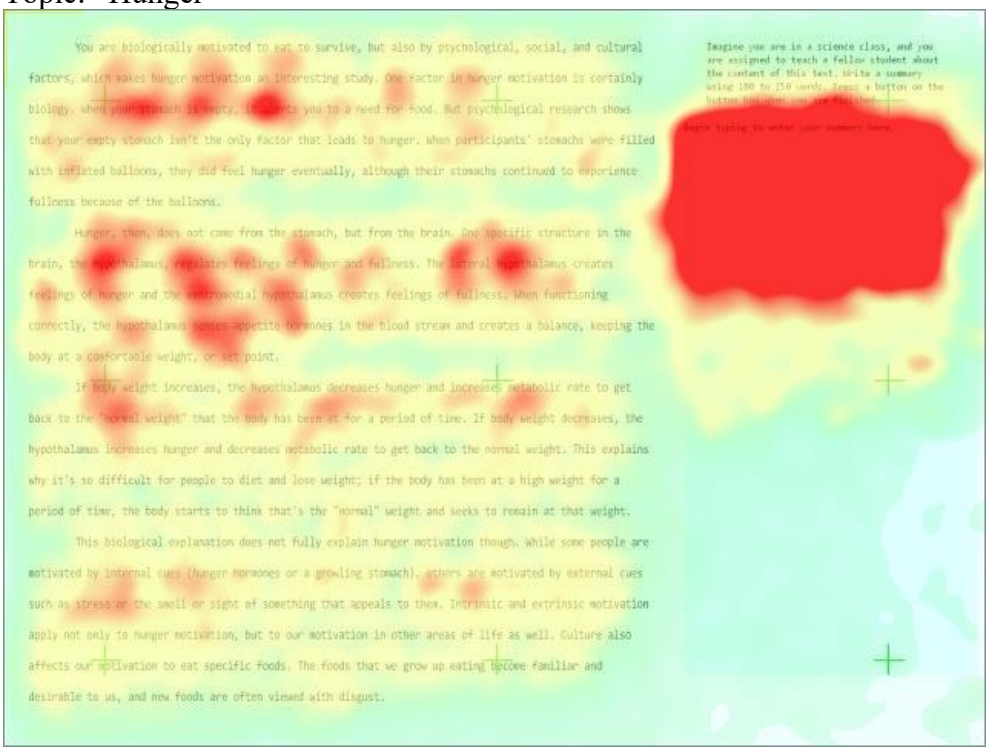
Topic: “Compound Microscope”



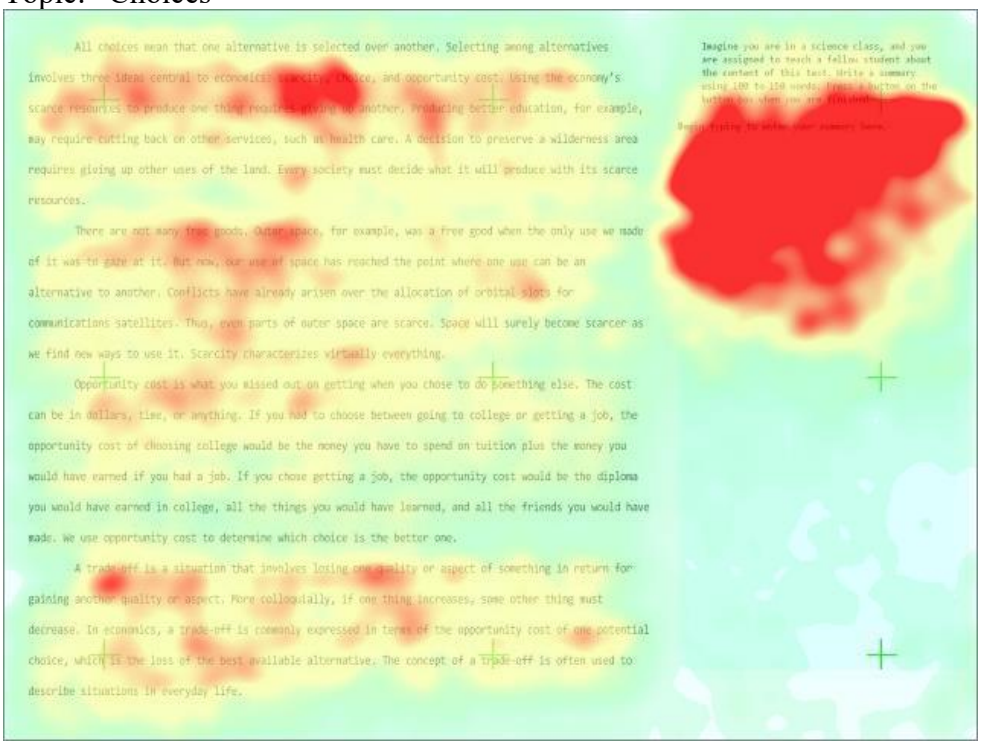
Topic: “Water”



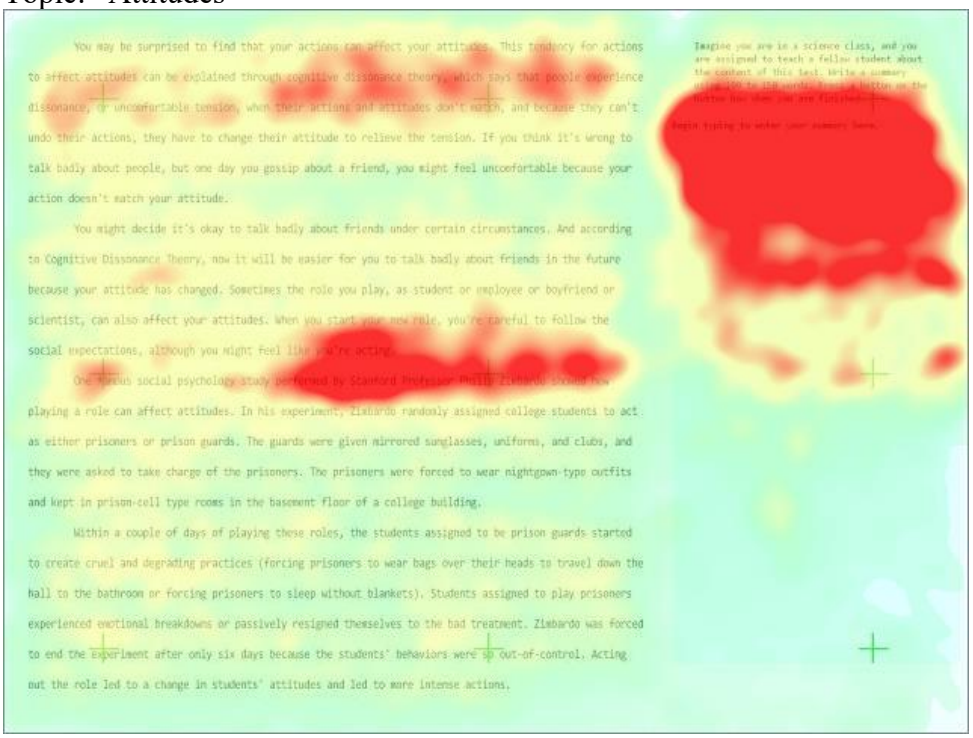
Topic: “Hunger”



Topic: “Choices”



Topic: “Attitudes”



Appendix H Eye-tracking descriptive statistics

This appendix presents descriptive statistics for eye-tracking metrics used in chapter 6. Table I.1 presents the means and standard deviations for each of the measures in total and across each reading task. These metrics were further analyzed across topics to see if some topics created unintended variance in eye behavior. These statistics give a good sense of the scope of each measure. They were compared in depth for significant differences in section 6.3. The eye-tracking metrics were further analyzed for normality by calculating skew and kurtosis. Similar to reaction times, there is a floor effect with real-time data, some positive skew is expected. After the removal of 12 outlier trials, skew and kurtosis values were calculated for each metric in each task. These results are presented in Table I.2. Although the distributions were different for metrics across the three tasks, no subset of data violated assumptions of skewness and kurtosis severely, with all skew measurements being greater than -1 and at most slightly over 1. No skew calculation was over 1.5. Kurtosis was within a satisfactory range (between 0 and 3) for all data.

Table H.1 Mean (SD) for eye-tracking measures

Task	Transitions	Mean Length of Saccade (pixels)	Fixations per word (Text)	Mean Fixation Duration (Text) (s)	Mean Fixation per word per Dwell (by line)
Cloze	96.61 (57.63)	83.6 (19.12)	3.263 (1.7)	0.260 (0.033)	0.201 (0.064)
MC	58.75 (30.93)	71.83 (16.23)	1.504 (0.703)	0.239 (0.034)	0.229 (0.098)
Summary	153.8 (100.03)	108.25 (44.06)	2.549 (1.225)	0.245 (0.028)	0.213 (0.075)
All	103.05 (79.13)	87.89 (32.89)	2.439 (1.463)	0.248 (0.033)	0.214 (0.081)

Task	Mean Fixation per word per Dwell (by paragraph)	Mean Fixation Duration (Task) (s)	Fixations per word (Task)	Average Rereading Duration (per line dwell) (s)	Average Rereading Duration (per paragraph dwell) (s)
Cloze	0.176 (0.102)	0.258 (0.07)	6.395 (3.872)	9.833 (6.398)	90.90 (49.20)
MC	0.143 (0.078)	0.145 (0.023)	1.84 (0.841)	3.132 (1.948)	32.91 (17.75)
Summary	0.106 (0.047)	0.368 (0.098)	3.948 (3.389)	6.702 (4.377)	63.10 (37.70)
All	0.141 (0.084)	0.257 (0.115)	4.061 (3.532)	6.555 (5.355)	74.098 (99.601)

Table H.2 Skew and Kurtosis for eye-tracking metrics

	Mean Length of Saccade		Transitions		Fixations per word (Text)		Mean Fixation Duration (Text)		Mean Fixation per Dwell (by line)	
Task	Skew	Kurt.	Skew	Kurt.	Skew	Kurt.	Skew	Kurt.	Skew	Kurt.
Cloze	1.081	1.151	0.894	0.407	1.321	2.148	0.509	0.917	0.979	1.183
MC	0.863	2.109	0.878	0.708	0.300	0.258	0.247	0.372	1.111	2.304
Summary	1.349	1.282	1.115	1.302	1.191	2.456	0.871	1.000	1.311	2.445

	Mean Fixation per Dwell (by paragraph)		Mean Fixation Duration (Task)		Fixations per word (Task)		Average Rereading Duration (per line)		Average Rereading Duration (per paragraph)	
Task	Skew	Kurt.	Skew	Kurt.	Skew	Kurt.	Skew	Kurt.	Skew	Kurt.
Cloze	1.447	1.810	-0.251	0.109	0.886	0.425	1.312	1.640	1.458	2.882
MC	1.009	1.600	-0.342	2.047	0.785	0.929	0.886	1.007	0.638	0.595
Summary	1.307	1.751	1.176	1.561	1.078	1.096	1.386	2.429	0.793	0.411

The metrics were also averaged for each task and topic condition and compared within tasks using one-way ANOVA to understand whether any topic effects were present. Means and standard deviations for each condition can be found in Table I.3. As 30 comparisons were made, the critical alpha was set at .002 using Bonferroni correction. Among the eye-tracking metrics, two significant differences for topic were identified. In the MC task, there was a significant topic effect for fixations per word in the text areas, $F(5,90) = 5.368$, $p < .002$, with topic 6 ("Attitudes") having a significantly smaller average fixation per word. There was also a significant topic effect for average duration of rereading per line dwell for the MC task, $F(5, 90) = 5.485$, $p < .002$, with topic 1 ("Biotechnology") eliciting more than the other topics. Due to the effect of topic in these cases, topic was further included as a random effect in subsequent analyses where applicable.

Table H.3 Mean (SD) for eye-tracking measures by topic

Task	Topic	Transitions	Mean Length of Saccade, pixels	Fixations per word (Text)	Mean Fixation Duration (Text) (s)	Mean Fixation per Dwell (by line)
Cloze	Biotechnology	111.71 (64.81)	86.63 (16.63)	3.31 (1.73)	0.26 (0.036)	0.16 (0.05)
	Compound Microscope	108.88 (28.3)	79.53 (16.09)	3.32 (1.11)	0.27 (0.036)	0.18 (0.04)
	Water	107.88 (55.47)	77.63 (14.35)	4.04 (2.26)	0.26 (0.037)	0.24 (0.09)
	Hunger	72.87 (36.31)	81.82 (21.2)	2.61 (1.06)	0.26 (0.026)	0.21 (0.05)
	Choices	111 (83.69)	80.18 (18.3)	3.37 (2.01)	0.26 (0.031)	0.21 (0.07)
	Attitudes	64.41 (39.16)	95.54 (23.57)	2.86 (1.46)	0.25 (0.026)	0.21 (0.05)
MC	Biotechnology	62.47 (32.78)	71.38 (8.24)	**2.02 (0.76)**	0.24 (0.029)	0.25 (0.11)
	Compound Microscope	65.53 (31.46)	73.04 (16.36)	1.77 (0.57)	0.24 (0.023)	0.2 (0.07)
	Water	58.82 (25.63)	83.75 (22.44)	1.53 (0.71)	0.24 (0.032)	0.24 (0.09)
	Hunger	53.71 (23.12)	69.61 (13.03)	1.30 (0.60)	0.24 (0.057)	0.22 (0.1)
	Choices	64.73 (36.36)	61.67 (7.24)	1.43 (0.45)	0.24 (0.022)	0.24 (0.06)
	Attitudes	47.94 (35.46)	70.34 (17.85)	**0.96 (0.64)**	0.23 (0.033)	0.22 (0.15)
Summary	Biotechnology	181.88 (101.5)	108.23 (32.39)	3.44 (1.75)	0.24 (0.033)	0.2 (0.06)
	Compound Microscope	145.94 (79.06)	113.35 (44.04)	2.06 (0.79)	0.24 (0.024)	0.2 (0.07)
	Water	160.18 (129.8)	103.52 (19.45)	2.52 (1.18)	0.24 (0.022)	0.22 (0.06)
	Hunger	147.47 (85.39)	112 (57.71)	2.48 (0.94)	0.24 (0.021)	0.21 (0.05)
	Choices	125.31 (65.78)	91.59 (34.26)	2.67 (1.19)	0.26 (0.022)	0.25 (0.11)
	Attitudes	159.88 (125.4)	120.11 (61.98)	2.11 (0.87)	0.25 (0.037)	0.21 (0.08)

Table H.3 (cont.)

Task	Topic	Mean Fixation per Dwell (by paragraph)	Mean Fixation Duration (Task) (s)	Fixations per word (Task)	Average Rereading Duration (per line dwell) (s)	Average Rereading Duration (per paragraph dwell)
Cloze	Biotechnology	0.15 (0.04)	0.27 (0.06)	7.44 (4.33)	9.85 (5.58)	98.81 (43.47)
	Microscope	0.21 (0.12)	0.27 (0.06)	6.72 (2.46)	8.37 (3.71)	98.72 (38.81)
	Water	0.18 (0.13)	0.25 (0.05)	7.19 (3.7)	13.3 (8.73)	114.71 (64.74)
	Hunger	0.17 (0.09)	0.26 (0.12)	4.86 (2.43)	7.93 (3.8)	73.40 (30.70)
	Choices	0.21 (0.13)	0.25 (0.08)	7.4 (5.58)	11.16 (7.57)	97.59 (58.59)
	Attitudes	0.13 (0.06)	0.24 (0.05)	4.55 (2.75)	8.01 (5.97)	60.19 (30.37)
MC	Biotechnology	0.15 (0.06)	0.14 (0.02)	2.01 (0.96)	**4.95 (2.39)**	45.81 (22.71)
	Microscope	0.14 (0.06)	0.15 (0.02)	2.44 (0.97)	2.7 (1.37)	39.83 (14.19)
	Water	0.14 (0.07)	0.14 (0.03)	1.57 (0.62)	3.04 (1.71)	31.02 (15.77)
	Hunger	0.15 (0.09)	0.15 (0.01)	1.49 (0.59)	2.78 (1.61)	27.54 (14.54)
	Choices	0.14 (0.05)	0.15 (0.01)	2 (0.68)	3.53 (1.48)	32.57 (12)
	Attitudes	0.14 (0.12)	0.14 (0.04)	1.55 (0.8)	1.85 (1.67)	20.65 (14.99)
Summary	Biotechnology	0.1 (0.04)	0.33 (0.05)	3.63 (1.93)	9.78 (7.31)	88.60 (63.30)
	Microscope	0.1 (0.04)	0.4 (0.12)	3.21 (1.91)	5.18 (2.29)	53.61 (24.27)
	Water	0.1 (0.05)	0.36 (0.08)	4.94 (4.57)	6.25 (2.92)	61.55 (27.52)
	Hunger	0.12 (0.05)	0.38 (0.1)	5.04 (5.32)	6.45 (3.84)	60.05 (31.32)
	Choices	0.11 (0.05)	0.42 (0.14)	3.16 (1.97)	7.23 (4.15)	66.09 (33.9)
	Attitudes	0.11 (0.06)	0.34 (0.06)	3.62 (2.73)	5.27 (2.38)	48.6 (19.81)

Note: Means and SDs marked in bold indicate a significant difference, $p < .002$, between the mean for that topic and others within the task condition.

Appendix I Linear mixed effect models predicting eye-tracking metrics by task

In each of the below linear mixed effects models (Tables J.1 through J.8), the models include task as a fixed effect with the cloze task as the baseline and text topic and individual participant as random effects. Model significance is presented below each model table. For specific pairwise differences between tasks for each metric, please refer back to chapter 6.

Table I.1 Predicting mean length of saccade

Task	*B*	SE	*t*	*p*	Marginal r_2	Conditional r_2
Intercept	84.033	2.878	29.198	< .001		
Multiple Choice	-12.326	3.231	-3.815	< .001		
Summary	23.316	3.225	7.229	< .001	0.217	0.516

Note: F(2,190) = 63.449, p < .001

Table I.2 Predicting number of transitions

Task	*B*	SE	*t*	*p*	Marginal r_2	Conditional r_2
Intercept	92.570	6.955	35.118	< .001		
Multiple Choice	-33.346	8.910	-3.743	< .001		
Summary	58.243	8.905	6.541	< .001	0.225	0.334

Note: F(2,187) = 54.515, p < .001

Table I.3 Predicting number of fixations per word

Task	*B*	SE	*t*	*p*	Marginal r_2	Conditional r_2
Intercept	3.151	0.153	20.544	< .001		
Multiple Choice	-1.625	0.133	-12.267	< .001		
Summary	-0.648	0.132	-4.901	< .001	0.254	0.531

Note: F(2,184) = 76.345, p < .001

Table I.4 Predicting mean text fixation duration

Task	*B*	SE	*t*	*p*	Marginal r_2	Conditional r_2
Intercept	0.260	0.003	87.469	< .001		
Multiple Choice	-0.018	0.002	-7.546	< .001		
Summary	-0.015	0.002	-6.311	< .001	0.065	0.723

Note: F(2,179) = 32.661, p < .001

Table I.5 Predicting mean fixation per line dwell

Task	*B*	SE	*t*	*p*	Marginal r_2	Conditional r_2
Intercept	0.198	0.009	21.105	< .001		
Multiple Choice	0.033	0.007	4.741	< .001		
Summary	0.011	0.007	1.598	0.112	0.034	0.598

Note: F(2,181) = 11.665, p < .001

Table I.6 Predicting mean fixation per paragraph dwell

Task	*B*	SE	*t*	*p*	Marginal r_2	Conditional r_2
Intercept	0.167	0.007	24.008	< .001		
Multiple Choice	-0.024	0.007	-3.288	0.001		
Summary	-0.062	0.007	-8.417	< .001	0.129	0.494

Note: F(2,185) = 36.037, p < .001

Table I.7 Predicting mean task fixation duration

Task	*B*	SE	*t*	*p*	Marginal r_2	Conditional r_2
Intercept	0.253	0.006	41.200	< .001		
Multiple Choice	-0.107	0.008	-14.140	< .001		
Summary	0.107	0.008	14.250	< .001	0.682	0.761

Note: F(2,190) = 406.04, p < .001

Table I.8 Predicting number of task fixations per word

Task	*B*	SE	*t*	*p*	Marginal r_2	Conditional r_2
Intercept	6.132	0.298	20.573	< .001		
Multiple Choice	-4.264	0.371	-11.507	< .001		
Summary	-2.313	0.370	-6.245	< .001	0.302	0.352

Note: $F(2,188) = 66.342$, $p < .001$

Appendix J Graphical comparison of eye-tracking metric means across reading tasks