Multiple-choice Reading Behaviors of ESL Students: An Eye-tracking Study

Juan Escalante*
Brigham Young University Hawaii, Laie, USA
(Corresponding author. Email: Juan.Escalante@byuh.edu)

Grant Eckstein
Troy L. Cox
Steven Luke
Brigham Young University, Provo, USA

Received: 22 July, 2022; Accepted: 15 December, 2022; Published: 24 January, 2023
https://doi.org/10.58304/tc.20230104

Abstract
Only recently has eye-tracking been used to investigate test-taker reading behavior, and results have been primarily used to confirm a range of cognitive tasks elicited by test items. This study explores test taker reading behavior for its own sake by describing how ESL readers of different proficiency levels behaviorally view multiple-choice passages and test items at different difficulty levels. Data were gathered from 51 students at three proficiency levels attending a university-sponsored intensive English program (IEP). Participants read eight validated reading comprehension items at varying difficulty levels while their eye movements were recorded on the passage, multiple-choice stem, correct answer, and distractors. Reading behavior demonstrated that language proficiency had a limited effect while passage difficulty had a stronger effect on reading behavior: participants gave less visual attention to the reading passage and correct answers within easier items and when they had higher language proficiency. The interaction of proficiency and item difficult on reading behavior is important in understanding how learners experience tests.

Keywords
Eye tracking, reading, multiple choice, test taking, proficiency, text difficulty

Introduction

Over the past several decades, researchers have widely investigated reading in a foreign language (Anderson, 2003; Aslanian-Nemagerdi, 1986; Park, 1997; Park, 2005). Major insights include that high-proficiency readers have a wider repertoire of strategies at their disposal, including the use of higher-order cognitive skills such as analyzing, reasoning, and elaborating in a second language. Conversely, less proficient readers tend to rely more heavily on bottom-up strategies and to focus mostly on literal meanings.

Problem-solving strategies, on the other hand, tend to be common in both proficient and less proficient learners. Theorists of native-English (L1) reading have been concerned with the overlap of reading and problem-solving for at least a century (Thorndike 1917, 1973–1974; Sternberg & Frensch, 2014), and various reading models, including bottom-up and top-down reading processing (Goodman, 1972; Gough, 1972), have ascribed problem-solving to top-down mechanisms. Additionally, Just and Carpenter (1987) as well as Rayner and Pollatsek (1989) have developed theories of pre- and post-lexical access and suggested that problem-solving can occur at both locations. In second-language literature, theorists have further noted that reading consists of strategic, purposeful, and interactive processes (Grabe, 2009), that it is a "constellation of interfaced capabilities" including "reasoning and inferencing" (Koda, 2004, p. 227), and that top-down processes are "directed by reader goals and expectations" (Grabe & Stoller, 2002, p. 32). Thus, it is reasonable to conceptualize reading as a strategic, problem-solving endeavor.

In test-taking situations, problem-solving is important because it can compensate for students' language proficiencies. For instance, Anderson (2003) demonstrated that both proficient and less proficient readers tended to rely heavily on problem-solving when engaged in online reading. Nevertheless, he also found that more and less proficient readers used the same strategies when taking reading comprehension tests, yet poorer readers used those strategies less successfully (Anderson, 1991). On the other hand, Allan (1992) found that "students are differentially skilled in test taking and that the scores of some learners may be influenced by skills which are not the focus of the test" (p. 101). These skills, collectively called *test-wiseness*, illustrate that problem-solving while reading can affect student performance. Thus, we can see that all learners problem-solve while reading, but they differ in how successfully they do it, and some learners may use problem-solving strategies to compensate for their proficiency limitations.

In spite of research showing the use of problem-solving in reading, less is known about learners' actual problem-solving behavior in test-taking situations. Current research so far has described test-taking reading processes through retrospective or reflective protocols and item analysis (Bax, 2013; Khalifa & Weir, 2009; Rupp, Ferne, & Choi, 2014; Weir, Hawkey, Green, & Devi, 2009). Meanwhile, less research has examined readers' test-taking reading behaviors in the moment through eye-tracking measures (Bax & Chan, 2016, 2019; Brunfaut & McCray).

As eye-tracking technology has developed, it has impacted our understanding of the processes elicited by reading in a foreign language. Eye-tracking has become a popular tool to investigate reading because readers are thought to attend just to those words which rest within their foveal vision, suggesting a link between reading behavior and cognitive processing (Rayner, 2009; Rayner, 1998). An early eye-tracking study of test-taking behavior revealed differences in how successful and unsuccessful test-takers viewed both reading texts and test questions (Bax, 2013). Other studies have shown that in a testing environment, readers tend to spend time searching for the correct answer rather than engaging with the text (Wang, Sabatini, O'Reilly, & Feng, 2017). However, eye-tracking research of test-based reading is still in its infancy and has not investigated differences in problem-solving among readers of different levels. Thus, the purpose of this research study is to examine how ESL learners of different proficiency levels differ in their reading processes as they read test items of varying difficulty levels.

Review of Literature

A general overview of reading

In order to examine problem-solving, it is expedient to first understand how reading has been defined over the years. This is no small task given the complex, multi-dimensional, and poorly-understood aspect of making meaning from text. Reading comprehension, as described by Bohn-Gettler and Kendeou (2014), involves encoding information, making inferences, and activating background knowledge. It also involves neurological and cognitive components (Hedgcock & Ferris, 2009). For instance, Wolfe (2007, p. 16) explained that "reading is a neuronally and intellectually circuitous act, enriched as much by unpredictable indirections of a reader's inferences and thoughts, as by the direct message to the eye from the text." Reading has also been defined as a process which is interactive, strategic, purposeful, evaluative, learning, and linguistic (Grabe, 2009). Grellet (1981) explains that understanding a written text is essentially extracting the necessary information as efficiently as possible. McCrudden, Magliano, and Schraw (2010) argue that not only is reading an intentional act but that it changes from one situation to another. For example, reading in an academic setting involves personal intentions as well as given intentions, thus readers bring personal knowledge and beliefs as well as externally provided instructions about what to do with the text. Others have defined reading as “a goal-directed activity in which the person intends to understand a text” (McCrudden & Schraw, 2007). Presumably, then, every student reading in an academic context should set goals to meet the task’s demands. These demands could be to summarize an article or to prepare to take the Test of English as a Foreign Language (TOEFL). In a rather comprehensive perspective, Koda (2004, p. 227) proposed that “adept reading is a constellation of interfaced capabilities, ranging from mechanical mappings to more sophisticated conceptual manipulations, such as reasoning and inferencing.”

Theorists have devised models of reading to make sense of the complex nature of reading. In an influential bottom-up model posited by Gough (1972), reading is thought of as the process by which a reader decodes letters or words, makes grapheme-to-phoneme associations, and constructs meaning from the text. Top-down models suggest that readers draw on background knowledge, engage with higher-level processes, and utilize hypothesis testing and checking (Goodman, 1972). Weir and Khalifa (2008, 2009) conceptualize reading as a multicomponential construct and have proposed a model of reading which can be broken down into strategies used to accomplish a given task. They indicate that the construct of reading needs to be described by observing the array of strategies readers use to construct meaning and complete the task. Based on these definitions and models of reading, it is evident that reading is a complex construct and that such complexity is likely to impact how ESL students read.

Models of problem-solving

Given this complex nature, it is not especially surprising that reading purposes effect the strategies which readers employ to comprehend a text. Narvaez et al. (1999), for instance, found that reading for entertainment versus study affected reader behavior in an inferencing task. Likewise, Zhang and Duke (2008) found that readers’ online reading strategies differed based on reading purpose (seeking information, acquiring general knowledge, or being entertained). Reading, therefore, is thought to be a dynamic, problem-solving process that changes depending on task specification.

Some of the best-known models of reading acknowledge this problem-solving component (Thorndike 1917, 1973–1974; Sternberg & Frensch, 2014; Goodman, 1972; Gough, 1972). The bottom-up model proposed by Gough (1972), which focuses on visual decoding of text, and Goodman’s top-down model that emphasizes background and schematic knowledge, both rely on readers cognitively resolving text-based problems (Sternberg & Frensch, 2014).

In a more recent model, Khalifa and Weir (2009) distinguish between local and global reading as well as careful and expeditious. *Local* reading takes place at the sentence level. It involves word recognition, lexical access, and synthetic parsing and establishing explicit propositional meaning at the phrase, clause, and sentence level. *Global* reading, on the other hand, deals with the comprehension of main ideas in a text. It entails going beyond the sentence level and understanding the connections between ideas in a text (Bax & Weir, 2012). C*areful* reading, which can happen at either global or local levels, entails extracting complete meaning from text. This type of reading is based on reading for comprehension, which is slow, careful, linear, and incremental. *Expeditious* reading, in contrast, involves quick, selective, and efficient reading in order to gain access to desired information in a text. This type of reading likely includes skimming, scanning, and search reading. In common with *careful* reading, *expeditious* reading can be present at the local or global level (Bax & Weir, 2012; Brunfaut & McCray, 2015; Khalifa & Weir, 2009). This model of reading has been used recently to investigate various issues associated with second-language testing, including cognitive validity of the International English Language Testing System (IELTS), cognitive processes employed by test takers, and the validity of multiple-choice reading tests (Bax & Weir, 2012; Brunfaut & McCray, 2015; Katalayi, & Sivasubramaniam, 2013; Weir, Hawkey, Green, & Devi, 2009).

Traditionally, researchers aiming to gain insight into readers' strategic, problem-solving/cognitive processes during reading have relied mainly on retrospective protocols (Bax & Weir, 2012; Brunfaut & McCray, 2015; Weir, Hawkey, Green, & Devi, 2009). However, recent studies have used eye-tracking technology to measure readers' actual reading behavior and thereby infer readers' cognitive processes while test-taking (Bax & Weir, 2012). Eye-tracking is considered to be superior to protocol tasks alone because it measures direct behavior rather than subjective participant memory or interpretation. Bax and Weir (2012), for instance, found that participants were only accurate in reporting their reading behavior 64% of the time, suggesting a large gap in what readers think they do versus what they actually do. Eye-tracking also reveals reading behavior which is automated or subconscious and therefore impossible for participants to report or be aware of.

Reading processes in test taking environments

Because eye-tracking methodologies are new to assessment research, few studies have examined the relationship between eye movement behavior and reading processes during test taking. The earliest study conducted by Bax and Weir (2012) collected eye-movement data from 15 high-performing students while completing five question sets from the Cambridge English: Advanced (CAE) test. The five items were chosen because they were thought to elicit all four combinations of careful/expeditious and local/global reading from the Kalifa and Weir (2009) model, and results confirmed this, leading the authors to conclude that the CAE text was cognitively valid. In a more intricate follow-up study, Bax (2013) examined the local-careful and local-expeditious reading behaviors of 38 Malaysian university students on two IELTS reading passages that were accompanied by 10 short-answer response questions. The aim of this research, like Bax and Weir's (2012), was to determine the cognitive validity of the assessment, and so the researcher compared the eye movements of successful and unsuccessful test takers. Results showed that in five of the 10 short-answer questions, successful readers read more expeditiously at the local level. Similarly, Wang et al. (2017) also examined local reading of 60 undergraduates on four tasks and found that participants read more expeditiously when preparing to answer multiple choice questions but more carefully in preparation for summary tasks.

In a well-designed mixed-methods study, Brunfaut and McCray (2015) examined 25 test takers' eye-tracking behaviors across four task difficulty levels (CEFR A1-B2) on the recently-developed Aptis test. Additionally, participants ranged in education level from pre-university to postgraduate and displayed different reading proficiency levels. Thus the study design allowed the researchers to examine reading behavior across task-difficulty and reader ability. Results showed that for 10 measures, readers' eye movements changed significantly by task difficulty: more difficult tasks generally required more visual and cognitive processing, as the researchers had hypothesized. The same was not entirely true for proficiency level since only four eye-tracking measures were significant; of those, all differences suggested that increased reading proficiency resulted in less visual and cognitive processing. Since this study was also conducted to investigate cognitive validity, no attempt was made to explore the interaction of task difficulty with readers' proficiency levels; furthermore, the tasks used were non-parallel, meaning their visual and interactive design differed one from another such that differences in reading processes could be attributed to the task design and not to the difficulty *per se*. This is unproblematic when seeking to design and validate a comprehensive reading test, but in terms of measuring reading differences across task and proficiency level, such an arrangement introduces unwanted variance.

To investigate global reading process, and to do so outside of the context of a validity study, Prichard and Atkins (2016) used eye-tracking measures to determine the previewing behaviors of 38 Japanese university students on a single reading task that was followed by an oral summary assessment. Results showed that participants largely failed to preview selectively at the global level, even when they self-ranked as high previewers. This finding comports with Bax (2013) and Brunfaut and McCray (2015) who found scant evidence of global expeditious reading but contrasts with Bax and Weir (2012) who did observe it. It should be noted, however, that Prichard and Atkins (2016) used a single reading passage as stimuli and participants were not grouped by proficiency; furthermore, the different test designs and time limits could have influenced reading behaviors and thus constrained the results.

Despite the insightful research conducted in these previous eye-tracking studies, many questions remain about the nature of assessment based-reading. Most eye-tracking studies involving test-taking have been cognitive validity studies rather than well-controlled investigations of reading behavior itself. Similarly, none of the previous studies examined interaction effects to illustrate how task and proficiency mutually contribute to reading behavior changes. This is partly because only one study, that of Brunfout and McCray (2015) utilized a mixed-task and mixed-proficiency design. The tasks they chose, however, were not parallel in their presentation, so differences in task difficulty are hard to interpret. Given these limitations and the need to better understand how students' reading processes change as task difficulty and reader proficiency levels increase, we designed an experimental eye-tracking study to investigate task-proficiency interactions on reading behavior of parallel multiple-choice reading items. We used the following research questions to guide our study:

1. How does reading proficiency and item difficulty affect students' initial and secondary reading times of multiple-choice reading passages, question stems, correct answer, and distractors?
2. How does reading proficiency and item difficulty affect students' look-back reading behavior of multiple-choice reading passages, question stems, correct answer, and distractors?

Methods

In order to investigate the ways in which language proficiency level and task difficulty affected the reading behaviors of ESL test-takers, we designed an eye-tracking experiment involving students in three language proficiency groups. They all read test-based passages with accompanying multiple-choice questions at three distinct difficulty levels. Below, we describe the methods used.

Participants

The participants in this study were 51 English-language learners from an intensive English program (IEP) associated with a large university in the western United States. Participants were recruited from pre-existing courses at three levels: 14 came from a Level 2 course, 18 from

Level 4, and the final group of 19 came from Level 6, with Level 6 being the highest level at the school. This IEP associates class levels with the American Council of the Teaching of Foreign Languages (ACTFL) proficiency guidelines. Level 2 roughly corresponds to intermediate-low in the ACTFL scale; Level 4 corresponds to intermediate-mid; and Level 6 corresponds to advanced-low. Students are placed in these levels based on performance exams across all language skills and conducted at the beginning of the semester (for new students) and end of previous semester (for returning students) by ACTFL-certified IEP instructors and administrators.

The 51 participants were nearly evenly distributed by gender: 26 males and 25 females. The age of the participants ranged from 18 to 47 years old (*M* = 26.7). In order to control for language, all participants were native Spanish speakers learning English as their second or additional language, and they came from the following ten countries: El Salvador, Honduras, Mexico, Argentina, Bolivia, Chile, Ecuador, Peru, Venezuela, and Spain. In addition, the participants' English-language learning backgrounds varied. For example, the length of time students had been enrolled in their current IEP ranged from 2 to 14 months in length. Similarly, the time they had spent studying English in their native countries varied, ranging from no instruction at all to as much as 13 years (1 hour per week, approximately), as shown in Table 1.

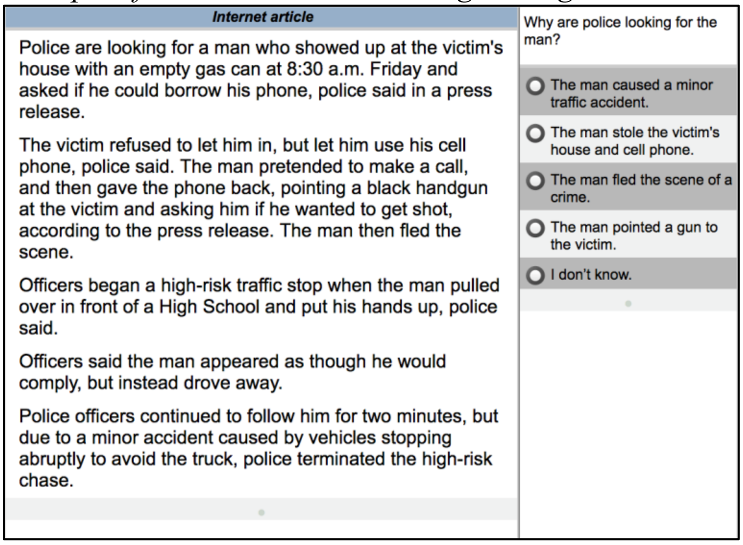
Table 1
Demographics of Participants

	Gender	Age (M)	Nationality	Months of English study outside of US (M)	Months of English study in current IEP (M)
Level 2	Male 8	25.7	Chile, Colombia, Mexico, Peru	8.7	2
	Female 6	24.7	Argentina, Ecuador, Honduras, Mexico, Peru	8.6	2
Level 4	Male 8	23	Argentina, Chile, Colombia, Ecuador, Peru	10	4.6
	Female 10	27.3	Chile, Ecuador, Guatemala, Honduras, Mexico, Peru, Spain, Venezuela	11	4.6
Level 6	Male 10	28.9	Bolivia, Chile, Colombia, Ecuador, El Salvador, Mexico, Peru	11.3	8.4
	Female 9	24.4	Bolivia, Chile, Colombia, Mexico, Peru	30.6	7

Materials

Eight reading comprehension test items were used in this study. They were developed by the Center of Language Studies at the study institution and had been previously administered with the item difficulty parameters aligning to expected proficiency levels, thus providing validity evidence for their use in a proficiency test. The selected reading passages represented three levels of ACTFL proficiency guidelines: intermediate, advanced, and superior. Reading passages at the intermediate level included a newspaper ad, a telephone note, and a sign for tourists, and the average word count was 31. At the advanced level, passages contained news articles that communicated straightforward, factual information as shown in Figure 1. The superior-level passages contained political commentary, analysis, and opinion editorials. The word count average between these two levels was similar, 155 for advanced and 158 for superior. Each passage was accompanied by a single multiple-choice question that aligned reader task with the author's purpose. All questions were designed to elicit inferential information where the correct answer had to be interpreted from general comprehension of the text. The question had five possible response options with one clearly correct answer, three distractors, and the fifth option labelled as "I don't know."

Figure 1
Example of Advanced-Level Reading Passage Used in this Study.



Apparatus
To measure eye movements, an SR Research Eyelink 1000 Plus eye-tracker with a 35 mm lens and a sampling rate of 1000 hz (1000 measurements per second) was used. As participants performed the tasks, they used a chin rest to reduce head movement and sat 65 cm away from a 21-inch computer monitor where the reading passages were displayed. Participants also used a computer mouse to select their responses to the multiple-choice questions. Prior to the experiment (and during, as necessary), a 9-point calibration was performed with every participant to ensure the collection of accurate data.

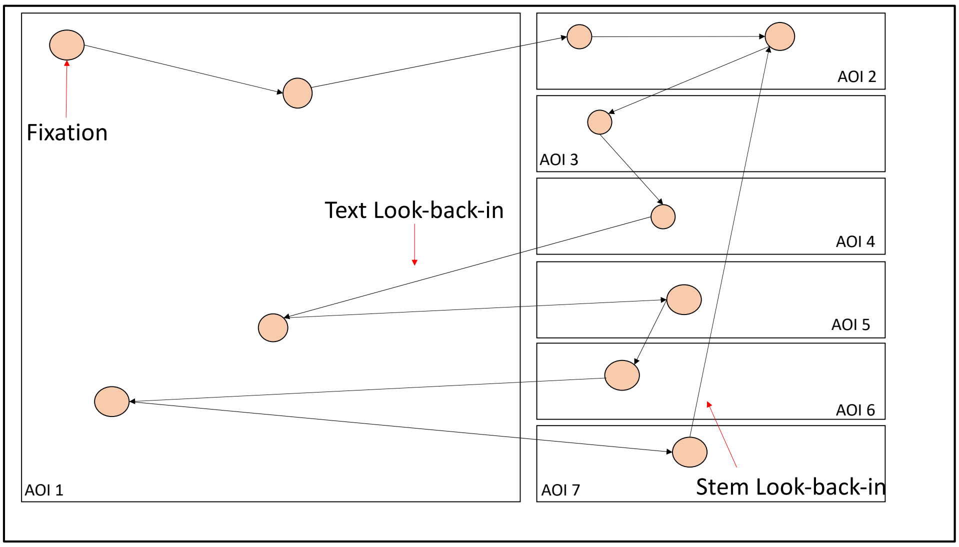
Procedures
To ensure uniformity in format, screenshots of the reading passages and their accompanying question were programmed into our eye-tracker. This uniformity was crucial since it allowed us to control for font size, location, and dimensions, and we were able to match the authentic online testing situation. We divided each passage into seven AOIs, one around the reading passage, one around the stem, and one for each of the five response options. These AOIs were essential to the collection of data because the eye-tracker records subjects' eye movement behavior for each preselected AOI, thus making it possible to collect eye-movement behaviors. We decided not to code each word as an AOI since the aim of the present study was to analyze reading behaviors at the global level rather than the word level.

Prior to the experiment, participants gave informed consent and were then positioned in front of the eye-tracking machine as described above. Participants were given verbal instructions and told that they would read nine short reading passages and that each passage would contain a multiple-choice question. They were instructed to read as naturally as possible. They were also encouraged to reread if necessary. In this study, we wanted to collect data that resembled natural test reading as closely as possible. Each reading passage and its accompanying question was presented together, one at a time, and in an increasing order of difficulty, intermediate level first, followed by advanced and superior. After the participants had read five of the nine passages, they were given the option to take a short break before continuing. On average, the experiment took approximately 30 minutes.

Data measurements

This study focused on examining the readers' interaction with the passage, stem, correct answer, and distractors in terms of first-pass dwell time, second-pass dwell time, and look-backs. First-pass dwell time (FPDT) refers to the total amount of time in milliseconds a reader spends on a specific area of interest (AOI) before moving on to a second AOI (Holmqvist, 2011). In word- and phrase-level studies, this measurement has been linked to early processing including letter recognition and word decoding (Liversedge, Paterson, & Pickering, 1998; Loftus & Mackworth, 1978). In passage-level research, it indicates the length of time a reader engages with a passage, question, or answer option before moving looking elsewhere. Second-pass dwell time (SPDT) refers to the sum of all fixations of an AOI during second-run reading and indicates re-reading for comprehension or to resolve confusion (Hyönä, Lorch, & Rinck, 2003). In passage-length research, SPDT indicates whether the passage, question, or answer options were consulted a second time. Look-backs, on the other hand, identify the number of times a reader returns to a specific AOI (Holmqvist, 2011). Clifton, Staub, and Rayner (2007) elaborate on look-backs by stating that this measurement can indicate effort by the reader to coordinate information and solve a problem. Figure 2 further elaborates on these definitions in an intentionally oversimplified illustration. In this figure, a hypothetical reading pattern is portrayed, and it can be noted that the first-pass dwell time in AOI 1 consists of two fixations. Likewise, the first-pass dwell time on AOI 2 only has two fixations. The second-pass dwell times for AOI 1 and 2 include one fixation each. The purpose of our research is not to indicate where in an AOI each fixation occurred, but rather to determine fixation frequency and relative interaction among the elements of the test question since we wanted to know how long participants spent dwelling on multiple-choice test items of varying difficulties including their passage, stems, correct answers, and distractors.

Figure 2
Scan Path Between AOIs.



Data analysis

This study was composed of three dependent variables (first-pass dwell time, second-pass dwell time, and look-backs) and two independent variables (test item difficulty and student proficiency). Several two-way ANOVAs were used in order to address the study's research questions, which examined student proficiency and item difficulty effects for the reading passage, the multiple-choice stem, correct answer, and distractors. Post-hoc Tukey-corrected pairwise comparisons were further used to analyze significant effects. Because there were three distractors on each item, we averaged eye-tracking measure for the three. We ran the analysis using Jamovi (The Jamovi project, 2019) which uses *R* (R Core Team, 2018; Fox & Weisberg, 2018). All data were log transformed prior to analysis.

Results

Overall, results showed main effects for item difficulty and student proficiency in terms of the reading passage, the correct answer, and the distractors. There were no effects for the stem, and no significant interaction effects for any measures. An overview of results is illustrated in Table 2 and a detailed analysis follows below.

Table 2

Main Effects for Student Proficiency and Item Difficulty

	FPDT		SPDT		Look-back	
	Proficiency	Difficulty	Proficiency	Difficulty	Proficiency	Difficulty
Passage	*			**		
Stem						
Correct	*	**		**		*
Distractor	*	**	*	**		*

Note: * $p < .05$; ** $p < .001$

Reading passage analysis

Results for the reading passage portion demonstrated few significant differences. Student proficiency was significant for FPDT in which level 2 readers spent about four seconds on average longer than level 6 readers ($p < .01$) indicating that higher proficiency readers spent less time in initial text reading (see Table 3), though the effect size was small (η^2p =.04). Item difficulty also moderated initial reading in that advanced and superior passages were read significantly longer than intermediate passages in SPDT with a strong effect size (*p*s < .01, η^2p =.15), though is to be expected since the more difficult passages were also longer.

Table 3

Eye-Tracking Measures for Reading Passage

			n[a]	M[b]	sd	Sem[c]	F	p	η^2p	Post-hoc
FPDT	Proficiency	Level 2	14	12520	10397	1688	3.08	0.049	0.041	Level 2 = Level 4
		Level 4	18	13003	18311	2492				Level 4 = Level 6
		Level 6	19	8864	12374	1639				Level 2 > Level 6
SPDT	Difficulty	Intermediate	51	4711	3635	509	12.71	< .001	0.15	Intermediate< Advanced
		Advanced	51	17530	19398	2716				Advanced = Superior
		Superior	51	17303	14477	2027				Intermediate < Superior

Note. = means no sig. difference; ≤ ≥ means sig. difference at .05; < > means sig. difference at .01
[a]For proficiency, *n* refers to the number of participants at each proficiency level while for difficulty, *n* refers to the total number of participants; [b]FPDT and SPDT mean is displayed in milliseconds; look back mean is calculated based on counts; [c]SEM is standard error mean.

4.2 Correct answer analysis

For the correct answer, text difficulty moderated FPDT insofar as Level 4 correct answers received longer dwell time than Level 6 by about half a second ($p < .01$) (see Table 4). The analysis also showed that the intermediate correct answers resulted in faster FPDT than superior correct answers by more than a half-second ($p < .01$). SPDT increased from intermediate to advanced to superior ($ps < .05$). Results show that superior level correct answers elicited more look backs than intermediate level correct answers ($p < .05$). This information confirms that more difficult items caused more interactions between the text and the answers.

Table 4
Eye-Tracking Measures for Correct Answer

			n^a	M^b	sd	Sem[c]	F	p	η^2p	Post-hoc
FPDT	Proficiency	Level 2	14	944	592	91.3	4.898	0.009	0.064	Level 2 = Level 4
		Level 4	18	1260	995	135.0				Level 4 > Level 6
		Level 6	19	806	530	70.2				Level 2 = Level 6
	Difficulty	Intermediate	51	671	365	51.1	8.595	< .001	0.107	Intermediate= Advanced
		Advanced	51	1039	951	133				Advanced = Superior
		Superior	51	1302	725	102				Intermediate > Superior
SPDT	Difficulty	Intermediate	51	694	335	46.9	16.722	< .001	0.188	Intermediate< Advanced
		Advanced	51	1121	570	79.8				Advanced ≤ Superior
		Superior	51	1645	1356	190.0				Intermediate < Superior
Look Back	Difficulty	Intermediate	51	1.31	0.79	0.112	3.537	0.032	0.047	Intermediate=Advanced
		Advanced	51	1.74	1.13	0.158				Advanced = Superior
		Superior	51	1.91	1.47	0.205				Intermediate ≤ Superior

Note. = means no sig. difference; ≤ ≥ means sig. difference at .05; < > means sig. difference at .01
[a]For proficiency, *n* refers to the number of participants at each proficiency level while for difficulty, *n* refers to the total number of participants; [b]FPDT and SPDT mean is displayed in milliseconds; look back mean is calculated based on counts; [c]SEM is standard error mean.

Distractor analysis

Results showed that proficiency affected FPDT for distractors. Level 4 distractors resulted in longer dwell times than level 6 ($p < .05$) (see Table 5). FPDT of distractors increased from intermediate to advanced to superior ($ps < .01$). SPDT showed that level 4 readers spent longer on distractors than level 6 ($p < .05$) and SPDT on distractors increased as item difficulty increased from intermediate to advanced to superior ($ps < .01$). Look backs to distractors were significant between intermediate and superior level texts ($p < .05$).

Table 5
Eye-Tracking Measures for Distractors

			n^a	M^b	sd	Sem[c]	F	p	η^2p	Post-hoc
FPDT	Proficiency	Level 2	14	953	524	80.9	3.56	0.031	0.047	Level 2 = Level 4
		Level 4	18	1026	591	80.4				Level 4 ≥ Level 6
		Level 6	19	839	443	58.6				Level 2 = Level 6
	Difficulty	Intermediate	51	521	179	25.1	69.13	< .001	0.49	Intermediate< Advanced
		Advanced	51	942	339	47.5				Advanced < Superior
		Superior	51	1346	584	81.7				Intermediate < Superior
SPDT	Proficiency	Level 2	14	1008	609	94.0	5.49	0.011	0.061	Level 2 = Level 4
		Level 4	18	1072	702	95.5				Level 4 ≥ Level 6
		Level 6	19	793	342	45.3				Level 2 = Level 6
	Difficulty	Intermediate	51	536	175	24.5	65.49	< .001	0.476	Intermediate< Advanced
		Advanced	51	1016	545	76.3				Advanced < Superior
		Superior	51	1299	611	85.5				Intermediate < Superior
Look Back	Difficulty	Intermediate	51	0.66	0.36	0.051	6.08	0.003	0.079	Intermediate= Advanced
		Advanced	51	0.86	0.66	0.093				Advanced = Superior
		Superior	51	1.14	0.89	0.126				Intermediate ≤ Superior

Note. = means no sig. difference; ≤ ≥ means sig. difference at .05; < > means sig. difference at .01
[a]For proficiency, *n* refers to the number of participants at each proficiency level while for difficulty, *n* refers to the total number of participants; [b]FPDT and SPDT mean is displayed in milliseconds; look back mean is calculated based on counts; [c]SEM is standard error mean.

Discussion

The purpose of our analysis was to examine test-taking reading behaviors on controlled multiple-choice test items that differed in difficulty and which were read by individuals at different language proficiency levels. As mentioned before, reading was measured using FPDT, SPDT, and look-backs. We expected that initial and secondary reading time of item passages would decrease with reader proficiency and increase with item difficulty based on evidence from Brunfaut and McCray (2015) which showed that visual attention decreased among higher-proficiency readers but increased with task difficulty. We also expected look-backs to change in similar ways since more proficient readers tend to be more strategic readers and therefore perform more look-backs (Bax, 2013). As expected, our results indeed showed that readers spent less time reading item passages as their language proficiency increased. Likewise, readers spent more time reading item passages of greater difficulty. However, this same pattern did not hold for selective reading insofar as we found no significant differences in passage look backs across item difficulty or reader proficiency.

A reasonable explanation of these findings is that more difficult passages used more complex lexical and syntactic structures which required greater processing time. Similarly, readers with limited proficiency were less likely to be familiar with these structures and therefore also required longer processing time. On the other hand, these results may simply be a product of the length of the passages given that in this study easier texts were shorter and used simpler words. However, texts at a lower ACTFL proficiency level necessarily use simpler language, and thus, we are unsurprised that text difficulty and language proficiency resulted in variances in textual reading. The fact that there were no changes in passage look-backs across item difficulty or proficiency levels suggests that item difficulty and reading proficiency did not modify how test-takers shuttled between question options and corresponding answer locations within the text; this is in contrast to Bax (2013) who found differences in the way readers located information based on reading proficiency. One interpretation could be that students across proficiency levels and task difficulty utilized similar test-taking strategies when interacting with question options and reading text. Wang et al (2017) found that test takers differed in their behavior when answering multiple choice items compared to summary tasks, thus little variation may have been found in the present study when only examining multiple-choice test-taking behavior. Given that the testing format for all items was the same (i.e., similar multiple-choice layout), it is likely that test takers were not obligated to shift their strategies among items. Or it could mean that readers found it unnecessary to shuttle much between question options and text, perhaps because the question options were easy to remember or answers were easy to find within the text. Stimulated recall data could provide more insights about this process.

Beyond the passage text, we were interested in how readers interacted with the correct answer. We expected that language proficiency would moderate reading time such that more proficient readers would need less time to look at and identify the correct answer. This bore out inasmuch as level 6 readers spent on average four-fifths of a second during FPDT while level 4 readers spent about 1.2 seconds, which was a significant difference. Since there were no significant differences in SPDT or look backs associated with reader proficiency, we speculate that readers only required a brief interaction with the correct answer to judge its acceptability. We also expected that item difficulty would increase correct answer reading times, and this certainly bore out with all reading measures increasing with item difficulty. In other words, readers spent more time visually attending to correct answers and returning to re-read them as item difficulty increased, which corroborates observations made by Brunfaut and McCray (2015). This seems to indicate two things. The first is that readers legitimately read and considered the correct answer from among the distractors, particularly as items became more difficult. But it also suggests that text difficulty, not language proficiency, caused test takers to be more selective—that is to move in and out of the correct answer to sample other components of the test item. In other words, strategic reading may be moderated more by item difficulty rather than reading proficiency. When students encounter more complex test items, they increased their strategic reading irrespective of their language proficiency. If this is indeed the case, it may be an indication that any kind of complex MC reading task inherently measures constructs of strategic reading in addition to time on text, at least when the reader has access to both the passage and question simultaneously.

The distractor analysis showed that proficiency of the readers as well as the difficulty of the items affected reading behavior of the distractors in confirmation of Brunfaut and McCray (2015). In particular, more proficient readers spent less time reading distractors, a finding that corresponds with typical reading expectations. However, proficiency was not a factor in look-backs, which may indicate that all readers employed similar distractor reading strategies. Nevertheless, item difficulty resulted in reading differences wherein more difficult items required greater careful reading and more selective distractor reading. This aligns with observations about the correct answer mentioned above and further indicates that readers seemed to shift their reading behaviors and make greater use of strategies when reading distractors on more difficult items.

The lack of interaction effects seems to indicate that proficiency level and text difficulty are independent factors. This means that individuals read differently based on their proficiency level and the text difficulty but not in combination. Furthermore, we saw evidence of reading changes with the passage itself, and these changes align with expectations of reading in that more proficient readers spent less time accessing the passage, and more difficult passages required greater reading times (Brunfaut & McCray, 2015). We saw the same pattern among correct choice and distractors, suggesting that readers in our study interacted with these elements in predictable ways. However, we also saw evidence of selective reading of the correct answer and distractors, which was not present when reading the passage itself. Combined, this seems to indicate that readers in our study accessed the correct answer and distractors strategically in an effort to problem solve by comparing these options with each other and with the passage. We interpret these results to mean that legitimate reading occurred with the passage and that strategic problem solving co-occurred because the text and the questions were displayed at the same time.

Conclusion

Evidence from this eye tracking study demonstrated that our participants read the textual component of multiple-choice questions in ways that align with previous protocol-based and eye-tracking studies of reading behavior during test-taking. That is, less proficient readers spent more time reading the passage than more proficient readers and more difficult texts resulted in longer reading times. Thus, one conclusion is that passage reading may be at least one component of multiple-choice test taking, even though skeptics might argue that multiple choice testing is merely an activity in problem solving. This has implications for TESOL teachers who may be curious if their students interact with the passage when completing MC tests. Our observation is that it does occur, or at least did occur in our data, and as a result, TESOL teachers may share this knowledge with their students and further encourage them to adopt passage reading strategies to make efficient and effective use of that reading time.

However, we also found evidence of problem solving as well. Test takers did not glance at the stem in any ways that significantly differentiated language proficiency or item difficulty; however, they did look at correct answers and distractors in ways that reflected natural reading (i.e., longer dwell times based on language proficiency and item difficulty) but they also showed significant measures of selective reading in terms of looking back to the passage, distractors, and correct answers, suggesting a high level of interaction among these features. We therefore conclude that multiple choice test taking may involve a dual process of both careful and selective reading in which readers are likely to carefully read the passage as part of a problem-solving activity and consult strategically with the correct answers and distractors, especially when encountering longer or more complex reading passages. It is important to mention that these observations are limited to the type of test format we used in this experiment, namely one in which passage texts and MC questions are displayed simultaneously.

While the current research was meant to evaluate readers' interactive behaviors on MC test items, several limitations make it difficult to generalize the findings. For instance, we collected eye-tracking data from just 51 participants, all of whom spoke Spanish as their first language, thus the findings cannot be generalized easily to a large and diverse population of ESL learners. It is possible that students from different language backgrounds, including native English speakers, may interact with reading passages, stems, and distractors in very different ways. In addition, there were a limited number of test items in this study, and we noted one grammatical error on a distractor in one test item, which may have skewed results slightly. Additionally, the intermediate passages were substantially shorter than those at other levels. Furthermore, only one MC question was used per test item which necessarily focused readers' attention on a relatively simplistic reading task and limited the kind of reading the task was likely to elicit. Our experimental design also failed to capture readers' comprehension scores due to a programming error, and we were unable to recover student selections. Thus, we were unable to distinguish the reading behaviors of more and less successful readers or otherwise match comprehension ability with reading proficiency.

These limitations should be controlled for in future studies by utilizing more test items across more task purposes and with standardized passage lengths. Stimulated recall procedures should also be included to get at readers' strategic approaches to test taking and triangulate eye-tracking findings. In the meantime, this study provides insights into the ways readers interact with various elements of a MC reading task. We found that passage difficulty and language proficiency effected how individuals interacted with MC tests. Specifically, participants gave less visual attention to the reading passage and correct answers within easier items and when they had higher language proficiency. These findings can benefit students, teachers, and test developers interested in understanding how proficiency and task difficulty affect reading behaviors in simulated test-taking environments.

References

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

Anderson, N. J. (2003). Scrolling, clicking, and reading English: Online reading strategies in a second/foreign language. *The Reading Matrix*, *3*(3), 1–15.

Anderson, N. J. (2017). Individual differences in strategy use in second language reading and testing. *The Modern Language Journal*, *75*(4), 460–472.

Ashby, J., Yang, J., Evans, K. H. C., & Rayner, K. (2012). Eye movements and the perceptual span in silent and oral reading. *Attention, Perception & Psychophysics, 74*(4), 634–640. Retrieved from https://search.proquest.com/docview/1323346254?accountid=4488

Aslanian-Nemagerdi, Y. (1986). *An investigation of the reading strategies of five proficient and five less proficient advanced ESL students* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (ProQuest Document ID No. 303464717)

Bahmani, R., & Mohammad, T. F. (2017). Effects of different text difficulty levels on EFL learners' foreign language reading anxiety and reading comprehension. *Reading in a Foreign Language, 29*(2), 185–202. Retrieved from https://search.proquest.com/docview/1979192802?accountid=4488

Baleghizadeh, S., & Yousefian, M. (2012). The relationship between test-takers' rate of strategy use and their reading comprehension performance. *Social Science Premium Collection, 48*(1), 73–115.

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

Bax, S., & Weir, C. (2012). Investigating figurative proficiency at different levels of second language writing. *Research Notes*, *47*(47), 14–28.

Bertram, R. (2011). Eye movements and morphological processing in reading. *The Mental Lexicon, 6*(1), 83–109. Retrieved from https://search.proquest.com/docview/881461334?accountid=4488

Beymer, D., Orton, P. Z., & Russell, D. M. (2007). An eye tracking study of how pictures influence online reading. In C. Baranauskas, P. Palanque, J. Abascal, & S. D. J. Barbosa (Eds.), *Human-Computer Interaction—INTERACT 2007* (pp. 456–460). Springer Berlin.

Blythe, H. I. (2014). Developmental changes in eye movements and visual information encoding associated with learning to read. *Current Directions in Psychological Science*, *23*(3), 201– 207. https://doi.org/10.1177/0963721414530145

Bohn-Gettler, C. M., & Kendeou, P. (2014). The interplay of reader goals, working memory, and text structure during reading. *Contemporary Educational Psychology*, *39*(3), 206–219. https://doi.org/10.1016/j.cedpsych.2014.05.003

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes whilst completing reading tasks: A mixed-method eye-tracking and stimulated recall study*. *ARAGs Research Reports Online. AR, 2015/001*. British Council.

Chiang, M. (2016). Effects of varying text difficulty levels on second language (L2) reading attitudes and reading comprehension. *Journal of Research in Reading*, *39*(4), 448–468. doi:10.1111/1467-9817.12049

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 341–371).

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

Cop, U., Drieghe, D., & Duyck, W. (2015). Eye movement patterns in natural reading: A comparison of monolingual and bilingual reading of a novel. *PLoS One, 10*(8). doi.org/10.1371/journal.pone.0134008

Fox, J., & Weisberg, S. (2018). *car: Companion to Applied Regression*. [R package]. Retrieved from https://cran.r-project.org/package=car.

Gibb, B. G. (1964). *Test-wiseness as secondary cue response*. Retrieved from ProQuest Dissertations & Theses Global (302180597). Retrieved from https://search.proquest.com/docview/302180597?accountid=4488

Goodman, K.S. (1972). The reading process: Theory and practice. In R. Hodges & E. Rudorf (Eds.), *Language and learning to read.* Houghton-Mifflin.

Gough, P.B. (1972). One second of reading. In J. Kavanagh & I. Mattingly (Eds.), *Language by ear and eye* (pp. 331–358). MIT Press.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice.* Cambridge UP. Grabe, W., & Stoller, F. (2002). *Teaching and researching reading.* Longman.

Grellet, F. (1981). *Developing reading skills.* Cambridge University Press.

Guo, Q., Kim, Y. –S. G., Yang, L., & Liu, L. (2016). Does previewing answer choice options improve performance on reading tests? *Reading and Writing*, *29*(4), 745–760. https://doi.org/10.1007/s11145-016-9626-z

Harasym, P. H., Price, P. G., Brant, R., Violato, C., & Lorscheider, F. L. (1992). Evaluation of negation in stems of multiple-choice items. *Evaluation & the Health Professions*, *15*(2), 198–220. https://doi.org/10.1177/016327879201500205

Hedgcock, J., & Ferris, D. R. (2009). *Teaching readers of English: Students, texts, and contexts*. Routledge.

Holmqvist, K. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Hyönä, J., Lorch, R. F., & Rinck, M. (2003) Eye movement measures to study global text processing. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research (*pp. 313–334). Elsevier Science.

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

Katalayi, G. B., & Sivasubramaniam, S. (2013). Careful reading versus expeditious reading: Investigating the construct validity of a multiple-choice reading test. *Theory and Practice in Language Studies*, *3*(6), 877–884. https://doi.org/10.4304/tpls.3.6.877-884

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

Koda, K. (2004). *Insights into second language reading: A cross-linguistic approach.* Cambridge University Press.

Lee, J. (2011, December). Second language reading topic familiarity and test score: Test-taking strategies for multiple-choice comprehension questions. UMI No. 3494064.

Liversedge, S. P., Paterson, K. B., & Pickering, M. J. (1998). Eye movements and measures of reading time. In G. M. Underwood (Ed.), *Eye guidance in reading and scene perception.* Elsevier.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive demands of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance, 4*(4), 565–572.

Macalister, J. (2010). Speed reading courses and their effect on reading authentic texts: A preliminary investigation. *Reading in a Foreign Language, 22*(1), 104–116. Retrieved from https://search.proquest.com/docview/1705667104?accountid=4488

McCrudden, M. T., & Schraw, G. (2007). Relevance and goal-focusing in text processing. *Educational Psychology Review*, *19*(2), 113–139. doi.org/10.1007/s10648-006-9010-7

McCrudden, M. T., Magliano, J. P., & Schraw, G. (2010). Exploring how relevance instructions affect personal reading intentions, reading goals and text processing: A mixed methods study. *Contemporary Educational Psychology*, *35*(4), 229–241. https://doi.org/10.1016/j.cedpsych.2009.12.001

Narvaez, D., Van Den Broek, P., Ruiz, A. B. (1999). The influence of reading purpose on inference generation and comprehension in reading. *Journal of Educational Psychology, 91*, 488–496.

Nation, P. (2009). Reading faster. *International Journal of English Studies, 9*(2), 131–144.

Park, G.-P. (1997). Language learning strategies and English proficiency in Korean university students. *Foreign Language Annals*, *30*(2), 211–221. https:/doi.org/10.1111/j.1944-9720.1997.tb02343.x

Park, H. (2005). *Second language reading: The interrelationships among text adjuncts, students' proficiency levels and reading strategies* (Order No. 3192443). Retrieved from ProQuest Dissertations & Theses Global. (304994776), https://search.proquest.com/docview/304994776?accountid=4488

R Core Team (2018). *R: A Language and environment for statistical computing*. [Computer software]. Retrieved from https://cran.r-project.org/.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422. Retrieved from https://search.proquest.com/docview/70115641?accountid=4488

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506. https://doi.org/10.1080/17470210902816461

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

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

Sabatini, J., Bruce, K., Steinberg, J., & Weeks, J. (2015). SARA reading components tests, RISE forms: Technical adequacy and test design. ETS Research Report Series, 2015(2), 1–20. doi:10.1002/ets2.12076

Sternberg, R. J., & Frensch, P. A. (2014). *Complex problem solving: Principles and mechanisms*. Psychology Press.

The jamovi project (2019). *jamovi*. (Version 1.0) [Computer Software]. Retrieved from https://www.jamovi.org.

Thorndike, E. L. (1917). Reading as reasoning: A study of mistakes in paragraph reading. *Journal of Educational Psychology, 8,* 323–332.

Thorndike, R. L. (1973–1974). Reading as reasoning. *Reading Research Quarterly, 2,* 135–147.

Tran, T. N. Y., & Nation, P. (2014). Reading speed improvement in a speed reading course and its effect on language memory span. *Electronic Journal of Foreign Language Teaching, 11*(1), 5–20.

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. doi.org/10.1080/10888438.2016.1276184

Weir, C. J., Hawkey, R., Green, A., & Devi, S. (2009). *The cognitive processes underlying the academic reading construct as measured by IELTS Authors*. *IELTS Research Reports Volume 9* (Vol. 9). Retrieved from http://www.ielts.org/researchers/research/volume_9.aspx

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

Wolf, M. (2007). *Proust and the squid: The story and science of the reading brain.* HarperCollins.

Yanagawa, K., & Green, A. (2008). To show or not to show: The effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *System*, *36*(1), 107–122. https://doi.org/10.1016/j.system.2007.12.003

Yang, P. (2000). *Effects of test-wiseness upon performance on the test of English as a foreign language* (Order No. NQ59700). Retrieved from ProQuest Dissertations & Theses Global, https://search.proquest.com/docview/304659398?accountid=4488

Zhang, S., & Duke, N. K. (2008). Strategies for internet reading with different reading purposes: A descriptive study of twelve good internet readers. *Journal of Literacy Research, 40*(1), 128-162.

Juan Escalante is an assistant professor of TESOL at Brigham Young University, Hawaii. His research interests include reading development using eye-tracking methods, teacher training, and program administration.

Grant Eckstein is the Associate Chair and assistant professor in the Linguistics Department at Brigham Young University. He studies second language reading and writing development, assessment, pedagogy, and curriculum design. He has published in venues such as *TESOL Quarterly* and *Research in the Teaching of English.*

Troy L. Cox (Associate Director, Center for Language Studies) is a faculty member in the Linguistics Department and has been involved with language teaching and research for over 20 years. He is a certified ACTFL oral proficiency tester and investigates second language proficiency, assessment, technology in teaching, self-assessment and objective measurement.

Steven Luke is an associate professor of Psychology and uses eye-tracking methods to investigate reading development and attention in first-, second-, and foreign-language contexts.