Language Testing in Asia

RESEARCH Open Access

Exploring the validity evidence of a high-stake, second language reading test: an eye-tracking study

Hyojung Lim

Correspondence: lim@kw.ac.kr
Department of English Language and Industry, Kwangwoon University, Hanwool 707, Kwangwoon-ro 20, Nowon-gu, Seoul, Korea

Abstract

The current study aims to explore the cognitive validity of the iBT TOEFL reading test by investigating test takers' eye movements on individual items. It is assumed that successful test takers would adopt the intended reading processes, the same types and levels of cognitive processes that they would use for real-world reading tasks. Forty-seven Chinese ESL students participated in the study, in which they took the TOEFL reading practice testlet on a computer, completed comprehension subskill tasks, and had stimulated recall interviews. Results showed that test takers tend to rely heavily on careful reading skills, while expeditious reading skills were rarely activated. The scope of reading was often restricted to the local level; learners hardly read more than a paragraph to answer questions. In some factual question items, successful readers were more efficient in reading and locating key information, whereas such group differences were not found in other items. Lastly, the gaze plots suggest that learners' eye movements manifest various interactions between comprehension subskills, primarily subject to bottom-up linguistic processing. The limitations and implications of learners' eye-tracking data for test validation will be further discussed.

Keywords: Cognitive validity, Eye-tracking, TOEFL, L2 reading, Comprehension subskills

Introduction

The concept of validity has been expanded and diversified for recent decades, resulting in variations (e.g., content validity, criterion validity, construct validity, cognitive validity, consequential validity); of fundamental interest to language testers would be whether a test measures what it is supposed to measure. In a proficiency test, the target construct, or the psychological reality defined by theoretical models, is often the object of discussion; a valid test should be able to tap the construct to be measured and thus test scores adequately reflect on the construct that test takers possess, or language ability.

To evaluate the construct validity of a language test, test validation has relied heavily on correlation-based statistical analysis. By approximating a test's factor structure, test developers intended to show that the test measures the latent construct of language ability (e.g., Sawaki, Stricker, & Oranje, 2009; Stricker & Rock, 2008). Otherwise, experimental evidence was examined through differential-groups and intervention studies (Brown, 2005). However, such statistical approaches based on test outcomes have gained much criticism because numbers do not convey conceptual information and test scores do not explain how test takers derived their answers (Weir, 2005). A number of scholars underlined the importance of understanding test takers' response behaviors and their cognition for the sake of test validation (Borsboom, 2005; Gorin, 2007). Without knowing test takers' thinking processes, any inferences and actions derived from test scores cannot be justified.

Along these lines, efforts have been made to inspect test takers' perspectives, perceptions, and cognitive processing. Cohen and Upton (2006) demonstrated test takers' reading and test-taking strategies on a TOEFL reading test by conducting retrospective interviews. Weir et al. (2009) examined the association between the skills measured by the International English Language Testing System (IELTS) reading test and what and how learners actually read in real-world academic settings through interviews and survey questionnaires. More recently, technology advances have enabled researchers to study cognitive validation in a wider variety of ways. Chan (2013) investigated ESL students' writing processes as they completed the writing section of the Pearson Test of English (PTE) academic test using a screen capture technique. Bax (2013) used eye-tracking technology to gain insight into the reading processes involved in the IELTS reading test. Arguments for cognitive validity necessitate the close scrutiny of test takers' mind. Cognitive validity means the extent to which test tasks elicit from test takers the same types and levels of cognitive processing as similar real-world tasks do or are expected to do (Bax and Weir 2012; Field 2011; Weir 2005).

Though language testers voiced the gravity of understanding learners' cognition, empirical studies are relatively scarce. Oftentimes, test validation has recourse to the analysis of test outcomes on the assumption that the test correctly exploited and thus measured the intended cognitive processes as well as the target construct. Of special interest to this study is therefore whether the TOEFL iBT reading test, a standardized second language (L2) academic reading test, in effect functions as designed by provoking the thinking processes inherent to academic reading under normal conditions. It is assumed that successful test takers would deploy the intended reading processes to arrive at a correct answer, and a valid test should be able to assess a wide range of cognitive processes that learners are to adopt in real-world academic settings.

Cognitive validity

While construct validity is a comprehensive and thus popular concept in language testing, cognitive validity appears relatively new. In terms of its conceptual and operational definitions, some investigated whether a test of scientific thinking or logical reasoning actually activates the processes it is supposed to measure (Baxter and Glaser 1998; Thelk and Hoole 2006). Others discussed how valid the test is as a predictor of real-life performance (Glaser 1981) and/or whether a test task elicits from test takers the same type and level of cognitive processing expected when performing real-world tasks (Bax and Weir 2012; Khalifa and Weir 2009; Weir et al. 2009). This sounds similar to

Bachman and Palmer's (1996) test-authenticity argument: "to justify the use of language tests, we need to be able to demonstrate that performance on language tests corresponds to language use in specific domains other than the language test itself." (p.23) Weir's (2005) context validity also accentuates the alignment between test tasks and real-world tasks in terms of linguistic and interlocutor demands. Context validity concerns "the extent to which the choice of tasks in a test is representative of the larger universe of tasks of which the test is assumed to be a sample" (p. 19). What makes cognitive validity distinct from the rest is that the former focuses on test takers' mental processes rather than on task characteristics. However closely test tasks simulate real-word tasks, they cannot be identical because of practical constraints. Thus, the focus of cognitive validity lies in the extent to which test takers' minds operate similarly to the way they do in real-world settings; cognitive validation in language testing is interested in learners' language skills, rather than linguistic knowledge, and their mental processes provoked by the test tasks rather than test scores.

The operational definition of cognitive validity has been treated, in math and science testing (Killen 2003) as the correspondence between the intended cognitive demands of the test task and the cognitive activity that the test actually invokes and/or the correspondence between the quality of the cognitive activity elicited by the test and test takers' performance scores. Hence, any positive correlation between the nature and scope of cognitive activity and students' test scores is often assumed to be evidence of cognitive validity (Baxter & Glaser, 1998); high scorers should be the ones who can deploy relevant cognitive resources and efficiently apply relevant thought processes. Similar points have been made by Field (2011). For an L2 speaking test to be cognitively valid, (a) the test tasks should be able to elicit from participants a set of processes that resemble those employed in real-word speaking events and (b) the relevant processes should be finely graded across levels in terms of the cognitive demands that they impose upon the participant. Field (2012) also suggested two ways to empirically examine cognitive validity in the IELTS listening test. The first involves comparing learners' language performance to experts' (e.g., native speakers or advanced learners). In such methods, native-like performance is treated as a primary criterion. The second involves comparing learners' mental processes in testing conditions to those activated in real-world settings. In this method, predictive validity serves as a primary criterion. Field (2013) concluded by summarizing three important features of cognitive validity: *similarity of processing, comprehensiveness*, and *calibration* (p. 80). The similarity of processing feature describes to what extent the cognitive processes elicited by a test are comparable to those that would be employed in a real-world setting, comprehensiveness refers to whether the range of processes elicited by a test is comprehensive enough to be representative of behavior in a real-world setting, and calibration addresses whether the cognitive demands imposed by a test are finely calibrated to reflect the level of the test. In the following section, empirical studies on the cognitive validity of language tests will be further reviewed in depth.

Cognitive validation: review of empirical studies

Discussions on the cognitive validity of language testing have increased among UK-based universities. Field (2012) attempted to evaluate the cognitive validity of the IELT S lecture-based listening test using verbal protocols. First, test takers' thinking processes were compared to those described by a conventional psycholinguistic model of first language listening. Second, learners' responses to listening under test conditions were compared to their responses under free "lecture" conditions. Third, the test scores were correlated to participants' listening performance in the real lecture setting. The retrospective verbal reports revealed that test takers frequently capitalized on test-wise strategies, such as exploiting cues provided by the wording or the ordering of items. The results showed that students focused on lexical matches rather than the main ideas of the lecture in question, indicating learners' relatively shallow processing under test conditions. Some students who performed badly on the listening test showed good comprehension in the non-test condition. Taken together, the author doubted the cognitive validity of the lecture-based question items and suggested that test authenticity could be improved by revising test formats and incorporating multi-modalities.

In research directly related to the current study, Bax and Wei (2012) conducted the first empirical research that investigated the cognitive validity of the *Cambridge English: Advanced (CAE)* reading test. Of 103 international students who participated in the study, the authors analyzed only six proficient learners' eye movements on five valid items, opting for a relatively conservative process of participant and item selection. For data analysis, the authors first developed nine predetermined criteria (p. 8) by themselves, identified reading types from test takers' eye movements accordingly, and calculated the agreement rates between the two analysts. It was observed that those who correctly answered the items also correctly adopted the reading processes the items were intended to activate. On one question, the students spent more time reading the question than the text, which the authors explicate indicates that readers had strongly constructed the meaning representation of the text by the time they had reached the question. This preliminary study took impressive strides in cognitive validation as it revealed the limitations of retrospective survey questionnaires; participants' actual reading behaviors diverged from their responses to the questionnaires. The findings also provide insight into the potential of eye-tracking technology as a research tool for language testing. In a more complete form, Bax (2013) compared successful test takers' reading processes to unsuccessful test takers' during the IELTS reading test. Of the original pool of 71 Malaysian undergraduate students, 38 students were randomly selected for eye recordings, and 20 of the eye-tracked students participated in the ensuing stimulated recall interviews. According to the results, (a) the unsuccessful readers read target words more slowly due to their limited vocabulary knowledge, (b) the successful readers were more efficient at locating key information and reading a text, and (c) unsuccessful readers sometimes read faster than successful readers but disregarded key information in the text. The author notes that the application of the eye-tracking methodology for test validation seems limited in this study given the question items analyzed predominantly relied on learners' word recognition skills.

More recently, Bax and Chan (2019) investigated eight Taiwanese learners' eye movements on the high-intermediate and advanced general English proficiency test (GEPT) Reading paper. As in the previous studies, the authors conducted a differential group study, while further inspecting the effect of item types (the cloze MC items, the MC reading comprehension items, and the cloze summary items). Their results showed that for some items successful readers did exhibit different cognitive processes from unsuccessful readers, such as spending more time on relevant areas in text and employing the expected cognitive processes. Stimulated recall interviews and students' self-report checklist as supplements substantiate the claim that different item types provoke different reading processes. Notably, one MC reading question that aims to elicit learners' intertextual activity failed to do so; students seemingly checked out the whole text, not to integrate information across paragraphs but to exert their skilled test-taking strategy (e.g., the elimination of distractors). Overall, there appears to be a real dearth in the literature of empirical studies on cognitive validation, despite its widely acknowledged value. The studies reviewed thus far pioneered the relevant research methodology and therefore represent valuable research. Having said that, the discussion of cognitive validity requires a clear understanding of a cognitive model of the skills to be measured; Field (2012) relied on psycholinguistic theories of speech perception and production when validating a speaking test. For a reading and writing test, Weir's socio-cognitive model has laid foundation for data analyses and interpretations (e.g., Bax and Weir 2012; Bax 2013; Bax and Chan 2019; Chan 2013). In this regard, the following section will provide an overview of the cognitive model of second language reading and its contribution to validating a reading test.

Cognitive model of second language (L2) reading

As in prior studies for test validation, the present study draws on Khalifa and Weir's (2009) cognitive processing model of real-life reading comprehension. The cognitive processing model originates from Weir's (2005) socio-cognitive framework, in which readers' internal processes are to interact with the context variables (e.g., task purposes, physical conditions of test administration, and linguistic features of reading passages) and the test taker characteristics (e.g., test takers' physiological and psychological attributes, reading experiences). Khalifa and Weir's reading model focuses on the internal processes, which comprise three major areas, executive processes, cognitive resources, and monitoring. The metacognitive activities, such as goal setting and goal monitoring, play a significant role in determining the types and levels of reading and thus the relative importance of the associated mental processes.

Under the cognitive processing model, reading is characterized as either careful or expeditious at the local or global level. Local comprehension takes place at the "level of decoding (word recognition, lexical access, and syntactic parsing) and establishing propositional meaning at the sentence and clausal level" (Khalifa & Weir, 2009, p. 45). Global comprehension covers larger scope of reading, such as understanding main ideas of a whole reading passage and recognizing text structures. Careful reading aims to extract complete meaning out of a given text. It is described as "slow, careful, linear, and incremental reading for comprehension" (Khalifa & Weir, 2009, p. 46). Note that careful reading is the default reading that most reading theories and eye-tracking studies take into account. Expeditious reading, on the other hand, refers to quick, selective, and efficient reading to locate specific information. The importance of reading speed for comprehension, alongside reading accuracy, has been highlighted by a number of scholars (Carver 1992; Overmann 1999; Weir et al. 2009); high correlations were reported between reading comprehension and reading fluency (Fuchs et al. 2001), Spanish-as-a-second-language learners in Judith (1995) counted scanning and skimming as the most necessary skills, and a causal relationship between word recognition skills and reading comprehension measures was reported in Verhoeven (2000). Khalifa and Weir's cognitive processing model identifies three kinds of expeditious reading skills, depending on a reading purpose; skimming, search reading, and scanning.

The core reading processing begins with bottom-up processes (word decoding, lexical access, and syntactic parsing). For the decoding processes to get going, strong vocabulary and grammar knowledge should be instantly available to a reader. Building on the automatized linguistic processing, meaning-making processing starts from relatively simple, explicit, and literal comprehension at the local level to complex, implicit, figurative reading comprehension at the global level. Beyond the textual comprehension at the clausal and sentential level, non-linguistic knowledge, such as readers' world knowledge, topical knowledge, and knowledge of text structure, should be established for use. The hierarchical nature of processing skills is entailed in Khalifa and Weir's (2009) review of cognitive processing across Cambridge ESOL levels. In the review, the A-level reading test is limited to assessing learners' bottom-up processing skills, ranging from word recognition to semantic proposition formation at the sentence level. From the B-level upwards, learners are expected to be able to exert their inferencing ability and build a meaning structure of a text. From the C-level upwards, learners start to be able to create a text level representation including main ideas and supporting details. Finally, C2 level students are distinguished from C1 level students in their ability of selecting, connecting, and organizing information from multiple texts. Adopted from Khalifa and Weir's, Bax (2013) clarified the levels of cognitive processing in reading tests (p. 3). Readers' cognitive processes are to start from simple activities, such as word matching, through more complex ones, such as building a mental model of a text or understanding the text function. The linguistic units involved in each step also evolve from a single word, through a sentence, to an entire text. Given the implicational and hierarchical nature of the cognitive processing in reading, making inference requires a certain level of automatization in lexical access and syntactic parsing. It is noteworthy that multiple parallel processing is also possible, meaning that any processes beneath can take place simultaneously (personal communication with Weir, 2014).

Khalifa and Weir took caution about relating the difficulty levels in processing to the hierarchical nature of processing in language tests. They reasoned that without considering linguistic factors, language tests cannot be solely determined by the complexity of cognitive processing, in that reading problems are often language problems for language learners (Jeon & Yamashita, 2014). With highly frequent vocabulary and simple grammar structures, inference questions can be easy, whereas with less frequent vocabulary and complex grammar structures, basic comprehension questions can be difficult. That said, Khalifa and Weir stated that readers' mental processing is "mediated by the contextual parameters of the text and task in hand" (p. 81). The contextual parameters involve the factors impacting on readers' cognition such as linguistic complexity and response formats.

The purpose of reading also comes into play in stimulating the cognitive processes with a differing level of emphasis. To scan a newspaper, the word recognition processing can be exclusively activated. For reading to write a research paper, readers' mental processing may go all the way up to the highest to create an intertextual representation. The metacognitive mechanisms (e.g., goal setting and monitoring) therefore mediate the different processing skills, while the goal setter comes in the first place to decide on a reading purpose and relevant cognitive processing, monitoring takes place throughout the whole reading process at different stages. Monitoring evaluates whether letters are decoded accurately and/or a writer's intention is understood successfully. The significant contributions of readers' metacognition to reading comprehension have been well documented by a number of empirical studies (e.g. Phakiti 2003; Purpura 1998; Song and Cheng 2009). All considered, Khalifa and Weir's (2009) cognitive processing model of reading can be an adequate reference to analyze the mental processing involved in the TOEFL iBT reading test.

Eye-tracking methodology

Eye-tracking technology is particularly useful in reading research because it provides moment-to-moment information that reflects readers' cognitive processing (Rayner, 2009). More recently, its application scope has been expanded; an increasing number of empirical studies examine learners' eye movements in second language studies, with a focus on word and grammar processing, vocabulary learning, multimodal information processing, language production and language assessment (Godfroid, 2020). Cognitive-control models provide a theoretical backing for the interpretation of eye-tracking data, assuming that cognitive processing determines readers' eye movements and eye fixations. For example, the eye-mind hypothesis suggests a strong association between eye movements and human mind (Just and Carpenter 1980; Morrison 1984; Pollatsek et al. 2006). Eye fixations and saccades are found to reveal what readers pay attention to, or which linguistic features receive readers' attention (Conklin et al. 2020; Reichle et al. 1998). Godfroid (2019) further specified that eye gaze represents one's focal attention or noticing, thereby emphasizing that skipped areas still receive peripheral attention. Several factors can affect the pattern of eye movements (Rayner 2009); longer fixations, shorter saccades, more regressions often stem from the complexity of linguistic features (e.g., word frequency or syntactic complexity) and/or readers' comprehension difficulty (e.g., learners' deficiency in language proficiency or reading skills). Namely, the duration and number of eye fixations account for how much cognitive effort readers make to comprehend the visual input. Readers' eye movements are also susceptible to a reading mode and a text genre; readers tend to fixate longer and saccades shorter when reading aloud (Laubrock and Kliegl 2015) and when reading non-fiction (Rayner et al. 2012). More relevant to the current study, the length of saccades and the angles between saccades can help to identify types of reading (Lemaire et al. 2011). The horizontality of saccades denotes the incidence of skimming and/or reading to memorize. Minimizing the distance to the current fixation turned out to be important for skimming, less important for memorizing, and not important for scanning.

Eye-tracking studies in reading research extensively investigated the effect of lexical properties on readers' eye movements, including word frequency, predictability, ambiguity, and the age at which a word is acquired (Juhasz & Rayner, 2006). However, little has been reported with regard to the impact of higher-order cognitive processing. Of the few empirical studies, Rayner, Chace, Slattery, and Ashby (2006) demonstrated how textual features affected readers' eye movements beyond the sentential level. In one experiment, the difficulty of a reading passage determined by a readability formula was correlated with average fixation duration, the number of fixations, and total reading time. As expected, more difficult text predicted longer fixation durations and more fixations. In a second experiment, readers' eye movements were compared when an anaphor was consistent or inconsistent with the antecedent across different distance levels. In near conditions, inconsistencies caused longer fixations and more regressions. As the distance between the anaphor and the antecedent increased, however, neither the number of regressions nor the fixation duration increased; thus, the inconsistency between the anaphor and the antecedent might not have been detected. Nevertheless, the second-pass reading time (i.e., the sum of all fixations in a region following the initial first-pass time, including the zero times when a region was not refixated) over the inconsistent antecedent was longer than that over the consistent one, possibly indicating that, apart from the distance effect, readers were more likely go back and reread the antecedent. Based on the findings, the authors concluded that eye movement data can adequately account for global reading difficulties and reading processing beyond the sentential level.

In regard to eye movement measures, Rayner et al. (2012) suggest using first-pass reading time (the time spent in a region before moving on or looking back), second-pass reading time (the duration of re-fixations), go-past time (the elapsed time from when a reader first enters a region until they move past it forward in the text), and total reading time in case the interest area is larger than a single word. However, the common measures for word reading have been first fixation duration (the duration of the first fixation on a word), single fixation duration (those cases where only a single fixation is made on a word), and gaze duration (the sum of all fixations on a word prior to moving to another word). More recently, Godfroid and Hui (2020) recommended running a principal component analysis of eye-tracking measures to reduce the number of measures to report and maximize the informativeness of statistics. Although eye movement measures should be selected a priori according to the intention of data analysis, researchers occasionally do not have options; some eye-tracking software programs such as Tobii Studio automatically generate pre-programmed metrics. Moreover, eye-tracking for the purpose of test validation is exploratory in nature, compared to conventional reading studies heavily focusing on lexical and syntactic processing. Following the eye-tracking studies published in the field of language assessment (e.g., Bax, 2013), the current study therefore consider six indices for eye-tracking analysis; the time to the first fixation (i.e., the time it takes before the eyes first fixate on the interest area), fixations before (i.e., the number of fixations on the stimuli before the eyes first fixate on the interest area), first fixation duration (i.e., the duration of the first fixation on the interest area), total fixation duration (i.e., the sum of the duration for all fixations within the interest area), total visit[1] duration (i.e., the duration of all visits within the interest area), and visit count (i.e., the number of visits within the interest area).

The study

To explore the cognitive validity of the iBT TOEFL reading test, successful readers' eye movements were compared with unsuccessful readers'. Such a group difference is assumed to reveal the degree of which each question item succeeds in eliciting the reading processes designed by test developers and thus make a valid distinction between good and poor L2 readers. With reference to the test specification of iBT TOEFL

[1] A visit is defined as the interval of time between the first fixation on the interest area and the next fixation outside the interest area

reading section, it is expected that those who scored high on the test would be proficient in *reading for information*, *reading for basic comprehension* and *reading to learn.* Note that the TOEFL iBT takes a reader-purpose approach to control task characteristics, rather than a task perspective or a processing perspective (Jamieson, Eignor, Grabe, & Kunnan, 2008). The reader-purpose perspective views reading "in terms of the coordinated application of knowledge and processes to a text or texts in the service of a goal or purpose" (Enright & Schedl, 2000). The TOEFL research team cast doubt on the processing approach for the sake of assessment by indicating that linguistic processes such as word recognition, fluency, and processing efficiency have limited value in explaining the hierarchy of language proficiency. However, Weir and his associates have shed light on the hierarchical nature of processing skills, specifying CEFR levels from the process perspective. An increasing number of language testers as well as reading researchers underscore the importance of assessing reading fluency at the tertiary level (Grabe 2010; Weir et al. 2009). The official guide to the TOEFL test (ETS 2009) also alludes to the importance of assessing reading fluency and rate; one characteristic of academic reading is effectively finding important information from text, thereby justifying the assessment of the reading-for-information type of reading. Hence, it seems legitimate to validate the TOEFL reading section with reference to cognitive processing model of reading, in which a purpose of reading still performs a predominant role. The research questions guiding the study are as follow.

Research questions

1. How do successful and unsuccessful learners differ in processing the first reading of the iBT TOEFL reading test?
2. How do successful and unsuccessful learners differ in processing the multiple-choice reading questions and the associated text?
   (a) How do they differ across question types (factual question, inference question, and vocabulary question)?
   (b) How do learners' comprehension subskills affect their eye movements?

Methods

Participants

A total of 47 Chinese ESL students from a midwestern university were invited to an eye-tracking lab to complete one reading testlet, extracted from the TOEFL iBT complete practice test volume 24, and several tasks that measure a range of comprehension subskills. The study included both undergraduate and graduate students with academic status varying from provisional to regular: 33 undergraduate students, 11 graduate students, and 3 provisional students. Thirty-one students were female. Note that out of 47, 46 students (98%) had sat a standardized English test before, either TOEFL or IELTS, 36 students (76%) had taken a language test preparation course, and 40 students (85%) reported a high degree of familiarity with the MC test format. Participants received $20 as compensation upon completing all required tasks. A summary of the demographic information of those who participated in the primary experiment is presented in Table 1.

Instruments

Reading test

One expository reading passage was selected from the TOEFL iBT complete practice test volume 24 (http://toeflpractice.ets.org/) with 14 associated reading comprehension questions. The testlet was first devised in HTML format and then installed in Tobii studio. The font was 12-point Times New Roman. As in the actual TOEFL iBT test, a text-only screen was presented, and the first question appeared on the next screen. Beginning with the second page, one question was presented on the left side of the screen, and the text remained on the right; by clicking the back and next buttons in the left upper-hand corner of the screen, a test taker was able to return to the previous questions or move forward to the next question. As in the actual TOEFL iBT test, text scrolling (up and down) was enabled. The results of item analysis and text analysis are summarized as below. Note that the last question (question 14) asks students to summarize the reading passage, allowing for partial credits. Hence, the weighted item was excluded from the item analysis. The rest was scored correct or incorrect. Only the items with item-total correlation values more than .25 (see Henning 1987) were submitted to the subsequent eye-tracking analysis (Table 2).

Comprehension subskill tasks

To tap into the sub-reading components of learners' reading ability, the following tasks were administered; the test of vocabulary knowledge (adopted from Schmitt et al. 2001), the test of grammar knowledge (adopted from Shiotsu 2003), lexical processing task (Lim and Godfroid 2015), sentence processing task (Lim and Godfroid 2015), and the automated symmetry span task (Redick et al. 2012). The vocabulary test had 90 questions in total, 30 for the 3000 level, the 5000 level, and the academic vocabulary, respectively. Participants were asked to choose one of six given words to match a definition. The grammar test consisted of 35 MC questions with four options. Both vocabulary and grammar test were untimed and paper-based.

In the lexical processing task, participants had to decide as quickly as possible whether a word presented on a computer screen referred to a living creature (e.g., a girl) or non-living artifact (e.g., a desk). Segalowitz and his associates (Segalowitz & de Almeida, 2002; Segalowitz & Frenkiel-Fishman, 2005) support the validity of the animacy judgment task, in that it tends to better provoke stronger semantic processing than a traditional lexical decision in which participants discriminate between real words and non-words. In the sentence processing task, participants were asked to choose a word that would come immediately after the beginning part of the sentence. For instance, the beginning of a sentence was provided (e.g., "After some time…") on the first screen, and on the next screen, two possible options followed (e.g., A. "works", B.

Table 1 Description of participants (*n* = 47)

	Mean	SD	Range
Age	20.93	2.63	19–31
Onset of learning (year)	9.85	2.79	5–16
Age of arrival (year)	19.72	2.84	16–28
Length of residence (year)	1.39	1.34	0.08–4
Number of test-taking experiences (TOEFL, IELTS, GRE, MSUELT)	3.02	1.35	1–8

Table 2 Item analysis of the 13 MC reading items (*N* = 47, *K* = 13)

	Mean	Std. deviation	Item discrimination	Corrected item-total correlation	Cronbach's alpha if item deleted
Item 1 (reference)	.64	.486	.65	.459	.636
Item 2 (vocabulary)	.32	.471	.26	.130	.686
Item 3 (inference)	.45	.503	.33	.154	.684
Item 4 (inference)	.64	.486	.34	.197	.677
Item 5 (vocabulary)	.79	.414	.41	.174	.678
Item 6 (factual)	.49	.505	.70	.496	.629
Item 7 (inference)	.45	.503	.39	.330	.657
Item 8 (factual)	.49	.505	.45	.256	.669
Item 9 (inference)	.66	.479	.70	.533	.624
Item 10 (vocabulary)	.81	.398	.35	.393	.650
Item 11 (factual)	.83	.380	.29	.196	.674
Item 12 (negative factual)	.85	.360	.35	.320	.660
Item 13 (cohesion)	.55	.503	.52	.352	.653

"she"). A participant was allowed to read the beginning at their own pace; however, he or she had to quickly select a word that best continued the phrase presented. The response time as well as accuracy rate were collected for both processing tasks. Lim and Godfroid (2015) provide empirical justification for the use of the sentence construction task, rather than the sentence verification task, to gage learners' syntactic processing skills. The response times as well as accuracy rates were recorded from both processing tasks; the coefficient of variance (CV) were calculated and submitted to the subsequent data analysis.

To measure participants' working memory capacity, we used the automated symmetry span task (Redick et al., 2012). In the task, participants made symmetry judgements regarding pictures while memorizing spatial locations. While the reading span task is more common in reading research studies, we intended to choose the symmetry task in an attempt to forestall the influence of learners' L2 proficiency. The reported Cronbach's alphas for the symmetry span task were .81 for partial scores and .73 for absolute scores (Engle, Tuholski, Laughlin, & Conway, 1999).

Stimulated recall interviews

Immediately after participants completed the MC reading test, a recall interview using gaze plots as stimuli was administered. The gaze plot displayed gaze data from one or several recordings as individual gaze points, fixations, and scan paths so that the order and length of fixations could be visualized. Because of time constraints, only three items that appeared in the latter part of the test were reviewed: one vocabulary question, one factual question, and one inference question. The stimulated recall sessions were conducted in English, audio-recorded, and transcribed. The purpose of the stimulated recall data was to confirm or supplement the identification of the types and levels of cognitive processes using eye movement data (e.g., Bax, 2013).

Data collection procedures

Every participant was invited to the eye-tracking lab, in which we first explained the entire procedure of the experiment to the individual and collected a consent form. Then, she took the reading testlet with 14 comprehension questions on a computer screen with the Tobii TX300 eye-tracker attached. Tobii TX300 records eye movements at a rate of 300 Hz (i.e., gaze data are sampled every 3.3 ms). A binocular camera was attached to the bottom of a 23" wide-screen TFT monitor. Unlike chinrest eye-trackers, the Tobii remote eye-tracker allows large head movements and is relatively unobtrusive and therefore ecological. The tracking distance between the eyes and the camera was held between 55 cm to 65 cm to optimize gaze accuracy. On average, gaze accuracy (i.e., a distance of 65 cm with 300 lux illumination) was 0.4°, whereas gaze precision 0.01° with a noise-reduction filter. To display the textual stimuli, collect eye movement data, and obtain eye movement measures (or metrics), Tobii Studio software was used. The velocity-threshold identification (I-VT) fixation classification algorithm was chosen to identify fixations and saccades. The I-VT filter was used to classify eye movement data based on the velocity of the directional shifts of the eye. As in Sereno and Rayner (2003), the minimum fixation duration was set at 100 ms, meaning that eye fixation was defined as the maintaining of the visual gaze for at least 100 ms. Before reading, the participant adjusted her sitting posture and performed a 9-point calibration. If a calibration point was missing or large errors occurred (e.g., a large difference between the gaze point calculated by the eye-tracker and the actual dot position), then the erroneous points were recalibrated. Twenty minutes were allowed for the reading test, during which the eye-tracker recorded participants' eyes. The weighted gaze samples turned out to be 90.03% on average, meaning that approximately 90% of the time, the eye-tracker successfully located and recorded learners' eyes. Immediately after the reading test, the stimulated recall interview and strategy questionnaires were completed. Then, I conducted the working memory test, the grammar test, and the lexical processing task in order. Participation was voluntary and $20 was given as compensation (Fig. 1).

Data analysis

As in Bax (2013), McCray and Brunfaut (2018), and Bax and Chan (2019), successful test takers' eye movements were compared with unsuccessful test takers'. Recall the assumption that successful readers are to deploy the expected cognitive processes on valid items, promptly adopt either expeditious or careful reading skills on demand, quickly locate relevant information in text, and pay more attention to key sentences or phrases, compared to unsuccessful readers.

To answer the first research question, the patterns of test takers' first reading, in which a text is presented without a question for the purpose of previewing, were examined. For this particular comparison, participants were grouped based on their total scores; the top 30% (N = 14) and the bottom 30% (N = 13) were tagged as high scorers and low scorers, respectively. Each group's total fixation duration on the entire reading passage was submitted to the t test. To further investigate any group differences in their first reading patterns, learners' eye fixation duration on each paragraph was submitted to the Kruskal-Wallis test. It was hoped to reveal which paragraph received more attention from readers.

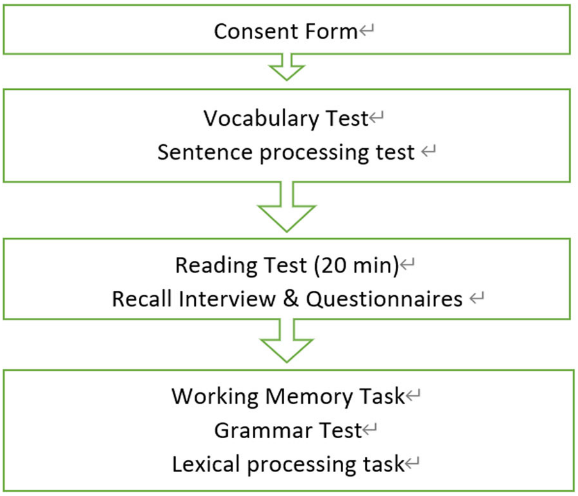


Fig. 1 Procedure of the study

To answer the second research question, the numerical eye-tracking data on each valid item were compared between those who got each question item correct (successful readers) and those who did not (unsuccessful readers). Note that the number of students for comparisons differed across items. First, we identified the areas of interest (AOIs) in the MC test as follows: (1) question stem, (2) response choice, (3) associated paragraph, and (4) key information in text. The key information, which provides cues to a key answer, consisted of one word to several sentences in the text (see Additional file 1). This information was determined by the author and double-checked by a non-linguist. The Mann-Whitney U test was conducted on the quantitative eye movement data for the AOIs (e.g., time to first fixation, fixations before, first fixation duration, total fixation duration, total visit duration, and visit count) when the normality and homogeneity of variance assumptions were not met. Otherwise, independent t tests were performed. To measure the test takers' attention and account for individual differences in reading speed, the reading time associated with the key areas was divided by the total reading time for the associated reading passage. In the "Results" section, however, we report only the criteria worth discussing with reference to statistical significance or meaningful insignificance. The gaze plots and interview data were examined to supplement the findings that emerged from the quantitative data analyses.

To explore the impact of test takers' sub-reading ability profile on their eye gaze, the collected data from the vocabulary and grammar test, the lexical and sentence processing task, and the working memory task were submitted to the t test for group comparisons. The gaze plots extracted from both groups were compared and examined qualitatively on a case-by-case basis.

Results

High scorers vs. low scorers: the first reading of the iBT TOEFL reading test

As shown in Table 3, high scorers demonstrated larger vocabulary and grammar knowledge and faster processing skills at the sentence level, compared to low scorers. However, the two groups did not seem much different from each other in processing the first reading part (Table 4). High scorers on average spent 76.99 s on the entire text, while low scorers spent 71.84 s. The difference was not statistically significant, however ($p = .91$). The results of the Kruskal-Wallis test showed that both high and low scoring groups equally spread out their attention over the six paragraphs ($\chi^2 = .728$, $p = .981$ for the high scoring group and $\chi^2 = 4.461$, $p = .485$ for the low scoring group); the first or the last paragraph did not necessarily receive more attention from the readers and thus the reading patterns did not differ across groups. It is also noteworthy that only the half of the entire test takers, regardless of their reading test scores, managed to complete the first reading, quickly rolling over their eyes from the first to the last paragraph and moving onto the first question. Taken together, evidence appears to be scant that either high scorers or low scorers used skimming skills properly on the first reading part.

Successful vs. unsuccessful readers: their reading processes across question types

Out of three vocabulary questions (question 2, 5, and 10), only one question (question 10) was found to be valid according to the earlier item analysis. Except for four participants (three successful readers and one unsuccessful reader), all checked out the associated paragraph to solve the question. The unsuccessful readers spent more time on the distractor A than the successful readers ($U(N_{\text{successful}} = 38, N_{\text{unsuccessful}} = 9) = 95.50$, $z = -2.05$, $p = .04$, $d = .62$ for total fixation duration, $U(N_{\text{successful}} = 38, N_{\text{unsuccessful}} = 9) = 97.50$, $z = -1.99$, $p = .05$, $d = .61$ for total visit duration, and $U(N_{\text{successful}} = 38, N_{\text{unsuccessful}} = 9) = 94.50$, $z = -2.12$, $p = .03$, $d = .63$ for visit count). The unsuccessful readers checked out the associated paragraph more frequently ($U\ (N_{\text{successful}} = 38, N_{\text{unsuccessful}} = 9) = 92.00$, $z = -2.14$, $p = .034$, $d = .66$ for visit count on paragraph 6). However, there was no group difference in terms of the amount of test takers' attention on the key sentence ($U\ (N_{\text{successful}} = 38, N_{\text{unsuccessful}} = 9) = 110.00$, $z = -.94$, p = .35, $d = .50$; Table 5).

Out of four, three factual information questions turned out to be valid: question 6, 8 and 12. In the question item 6, it took more time for the unsuccessful group to finish reading the question stem ($U(N_{\text{successful}} = 23, N_{\text{unsuccessful}} = 24) = 174.50$, $z = -2.16$, $p = .031$, $d = .66$). In terms of text reading, the successful group paid more attention on the key phrases ($U(N_{\text{successful}} = 23, N_{\text{unsuccessful}} = 24) = 103$, $z = -3.07$, $p = .002$, $d = 1.27$), while they read the associated text faster than the unsuccessful group ($U(N_{\text{successful}} = 23, N_{\text{unsuccessful}} = 24) = 172$, $z = -2.21$, $p = .027$, $d = .68$). The successful group likely picked up the key words/phrases in the text faster and revisited them more frequently, compared to the unsuccessful group; the former caught the key phrase *no*

Table 3 Total fixation duration on the first reading

	High scorers (N = 14)				Low scorers (N = 14)			
	Mean	SD	Kruskal-Wallis	N	Mean	SD	Kruskal-Wallis	N
1st paragraph	8.15	8.38	$\chi^2 = .728$, $p = .981$	14	11.26	11.91	$\chi^2 = 4.461$, $p = .485$	13
2nd paragraph	13.20	16.31		12	15.60	24.19		10
3rd paragraph	13.09	18.56		10	7.91	13.32		8
4th paragraph	14.47	13.51		7	15.29	22.87		5
5th paragraph	12.32	14.26		8	13.43	23.41		4
6th paragraph	15.76	17.32		7	8.35	14.01		5
SUM	76.99				71.84			

Table 4 The sub-reading abilities of each group

	High scorers (***N*** = 14)	Low scorers (***N*** = 14)	***t*** test
Vocabulary knowledge	79.85 (7.41)	58.07 (12.71)	$t = 5.54$ ($p < .01$), $d = 2.09$
Grammar knowledge	30 (2.80)	19.57 (4.62)	$t = 7.22$ ($p < .01$), $d = 2.73$
Lexical processing skills	0.26 (0.08)	0.30 (0.06)	$t = -1.63$ ($p = .12$), $d = .57$
Sentence processing skills	0.40 (0.11)	0.62 (0.13)	$t = -4.77$ ($p < .01$), $d = 1.83$
Working memory capacity	23.57 (7.95)	20.21 (7.87)	$t = 1.12$ ($p = .27$), $d = .43$

stream and *steady water* more quickly than the latter. In case of the key phrase *steady water*, the difference was statistically significant ($U(N_{\text{successful}} = 23, N_{\text{unsuccessful}} = 24) = 48.00$, $z = -2.49$, $p = .013$, $d = 2.00$, for time to first fixation). The successful group's fixation count and visit count for the phrase *no stream* was significantly higher than the unsuccessful group ($U(N_{\text{successful}} = 23, N_{\text{unsuccessful}} = 24) = .05$, $z = -2.01$, $p = .045$, $d = 3.32$, and $U(N_{\text{successful}} = 23, N_{\text{unsuccessful}} = 24) = 44.50$, $z = -2.10$, $p = .036$, $d = 2.07$, respectively), possibly indicating that the successful group was actively matching the phrase *no stream* in the key option to that in the given text. A likely explanation for the different statistical results of the key words is that in the case of *no stream*, the exactly same wording appeared in the option and the text, whereas *steady water* in the text was replaced with *constant water* in the option. Hence, the key word *no stream* could have encouraged high scorers' word-matching behavior, presumably explaining a higher number of visits on the phrase *no stream,* while the other key phrase *steady water* did not yield the same effect (Table 6).

In item 8, no group differences were found in terms of text reading. Compared to the unsuccessful readers, the successful readers spent more time reading the text and the key sentence. However, the difference was not statistically significant. Considering the reading time spent on the paragraph and the key area, the successful readers appeared to give more attention on the key sentence. However, the difference was not statistically significant, either (U ($N_{\text{successful}} = 23$, $N_{\text{unsuccessful}} = 24$) = 197.50, $z = -1.47$, $p = .141$, $d = .50$). When it comes to their reading patterns on the question and options, the successful readers spent more time on the key option C and the distractor D ($U(N_{\text{successful}} = 23, N_{\text{unsuccessful}} = 24) = 156.50$, $z = -2.54$, $p = .011$, $d = .80$ for total fixation duration on option C and $U(N_{\text{successful}} = 23, N_{\text{unsuccessful}} = 24) = 169.50$, $z = -2.27$, $p = .023$, $d =$

Table 5 Eye-tracking statistical results for item 10

	Total fixation duration on option A	Total visit duration on option A	Visit count on option A	Visit count on paragraph 6	Visit count on the key sentence	Attention on the key sentence
Incorrect (*N* = 9)	1.28 (.87)	1.28 (.87)	2.78 (1.56)	11.89 (9.32)	5.67 (5.61)	.30 (.25)
Correct (*N* = 38)	.70 (.69)	.72 (.71)	1.66 (1.17)	6.55 (8.64)	3.79 (3.33)	.42 (.31)
Mann-Whitney *U*	95.50	97.50	94.50	92.00	138.00	110.00
Z	− 2.05	− 1.99	− 2.12	− 2.14	− .90	− .94
Sig. (2-tailed)	.041	.046	.034	.032	.369	.349
Cohen's *d*	.62	.61	.63	.66	.26	.50

Table 6 Eye-tracking statistics for item 6

	Total fixation duration on stem	Total visit duration on stem	Attention on the key phrases	Total fixation duration on paragraph 3	Time to first fixation on *steady water*	Time to first fixation on *no stream*	Visit count on *steady water*	Visit count on *no stream*	Fixation count on *steady water*	Fixation count on *no stream*
Incorrect (N = 24)	5.70 (4.08)	7.13 (5.22)	.09 (.08)	39.08 (22.63)	57.48 (25.71)	64.16 (22.09)	4.13 (2.59)	2.31 (1.38)	6.20 (4.13)	2.54 (1.61)
Correct (N = 23)	3.52 (1.98)	4.21 (2.60)	.22 (.15)	24.66 (19.76)	38.22 (31.04)	47.46 (32.47)	5.79 (3.02)	3.85 (2.04)	9.14 (5.11)	4.23 (2.32)
Mann-Whitney U	174.50	161.00	103.00	172.00	48.00	50.00	70.50	44.50	.12	.05
Z	− 2.16	− 2.45	− 3.07	− 2.21	− 2.49	− 1.77	− 1.52	− 2.10	− 1.55	− 2.01
Sig. (2-tailed)	.031	.014	.002	.027	.013	.077	.129	.036	.120	.045
Cohen's d	.66	.76	1.27	.68	2.00	1.97	1.66	2.07	3.32	3.32

.70 for total fixation duration on option D). In particular, the successful readers more repeatedly checked out the key option C (U ($N_{\text{successful}}$ = 23, $N_{\text{unsuccessful}}$ = 24) = 144.50, z = − 2.82, p = .005, d = .89 for visit count on option C), compared to the unsuccessful readers (Table 7).

In item 12, which asks readers to pick false factual information, both groups were similar in reading the given paragraph. The successful readers spent more time reading the text and the key areas, compared to the unsuccessful readers, which was not statistically significant (U ($N_{\text{successful}}$ = 40, $N_{\text{unsuccessful}}$ = 7) = 133.00, z = − .21, p = .834, d = .061 for total fixation duration on paragraph 6). As regards the question and options, the unsuccessful readers likely spent more time on the distractor A and B (U($N_{\text{successful}}$ = 40, $N_{\text{unsuccessful}}$ = 7) = 56.00, z = − 2.51, p = .012, d = .79 for total fixation duration on option A and U($N_{\text{successful}}$ = 40, $N_{\text{unsuccessful}}$ = 7) = 73.00, z = − 2.00, p = .045, d = .61 for total fixation duration on option B, respectively), while the key was option D (Table 8).

Out of three inference questions, two turned out to be valid, question 7 and 9. In item 7, the unsuccessful readers looked at the question itself more frequently than the successful readers (U($N_{\text{successful}}$ = 21, $N_{\text{unsuccessful}}$ = 26) = 179.50, z = − 2.01, p = .045, d = .61 for visit count on stem), possibly indicating their difficulty in registering what is being asked. One of the unsuccessful readers on item 7 recalled during the interview that "the thing is I can't really concentrate on the first or the second time I read something. Like I am just reading it but not actually, like reading it, just see it. So, I have to…and I saw a term I am not familiar with, like, molecule". The unsuccessful readers also checked out the distractor A more frequently and for a longer period of time (U($N_{\text{successful}}$ = 21, $N_{\text{unsuccessful}}$ = 26) = 151.50, z = − 2.62, p = .009, d = .82 for visit count on option A, U($N_{\text{successful}}$ = 21, $N_{\text{unsuccessful}}$ = 26) =128.00, z = − 3.10, p = .002, d = 1.015 for total fixation duration on option A). Unexpectedly, the unsuccessful readers appeared to give more attention to the key area in the text, which shows a substantial trend towards significance (U ($N_{\text{successful}}$ = 21, $N_{\text{unsuccessful}}$ = 26) = 188.00, z = − 1.82, p = .069, d = 0.55). This was the opposite pattern observed in item 6, where it was the successful readers who paid more attention to key areas (Table 9).

In item 9, the successful readers looked at the key option A for longer (U($N_{\text{successful}}$ = 31, $N_{\text{unsuccessful}}$ = 16) = 159, z = − 2.00, p = .045, d = .61 for fixation count, U($N_{\text{successful}}$ = 31, $N_{\text{unsuccessful}}$ = 16) = 161.00, z = − 1.95, p = 0.51, d = .59 for total visit duration), whereas the unsuccessful readers stayed on the distractor D more (U($N_{\text{successful}}$ = 31, $N_{\text{unsuccessful}}$ = 16) = 151.50, z = − 2.17, p = .030, d = .67 for fixation count, U($N_{\text{successful}}$ = 31, $N_{\text{unsuccessful}}$ = 16) = 133.00, z = − 2.58, p = .010, d = .81 for total fixation duration). It took almost double the time for the unsuccessful readers to read the associated text, which almost approached significance (U ($N_{\text{successful}}$ = 31, $N_{\text{unsuccessful}}$ = 16) = 166.00, z = − 1.84, p = .066, d = .56). However, the amount of attention on the key sentence was not dissimilar between the groups (Table 10).

With regard to other item types, in item 1, the reference question, the unsuccessful readers looked at the distractor B longer (U($N_{\text{successful}}$ = 30, $N_{\text{unsuccessful}}$ = 17) = 102.00, z = − 3.39, p = .001, d = 1.14), whereas the successful readers stayed on the key option D longer (U($N_{\text{successful}}$ = 30, $N_{\text{unsuccessful}}$ = 17) = 129.50, z = − 2.78, p = .005, d = .89). It took more time for the unsuccessful readers to finish reading the relevant text than the successful readers (U ($N_{\text{successful}}$ = 30, $N_{\text{unsuccessful}}$ = 17) = 150.00, z = − 2.33, p = .020, d = .72). With regard to the amount of their attention on key phrases, however, no group

Table 7 Eye-tracking statistics for item 8

	Total fixation duration on option C (key)	Total fixation duration on option D	Fixation count on option C (key)	Fixation count on option D	Visit count on option C (key)	Total fixation duration on paragraph 5	Total fixation duration on the key sentence	Attention on the key sentence
Incorrect (*N* = 24)	2.70 (1.99)	1.68 (1.24)	8.75 (6.18)	5.42 (3.87)	3.46 (2.13)	29.76 (13.95)	1.56 (1.56)	.05 (.04)
Correct (*N* = 23)	4.56 (2.68)	2.57 (1.40)	14.57 (7.87)	8.26 (4.51)	6.00 (3.13)	33.80 (23.30)	2.39 (2.19)	.07 (.04)
Mann-Whitney *U*	156.50	169.50	157.50	177.50	144.50	269.00	213.00	197.50
Z	− 2.54	− 2.27	− 2.53	− 2.11	− 2.82	− .15	− 1.34	− 1.47
Sig. (2-tailed)	.011	.023	.012	.035	.005	.882	.179	.141
Cohen's *d*	.80	.70	.79	.64	.89	.04	.40	.50

differences were found. In item 13, the cohesion question, no group differences were detected, either in question reading or in text reading (Table 11).

To recap, the eye-tracking data seem to corroborate that successful readers are more efficient in reading a text and a question than unsuccessful readers; successful readers were faster in reading either a text or a stem in question 1, 6, 7, 9, and 10. The eye movements on item 6 best demonstrated successful readers' efficiency in locating important information. One unsuccessful reader (ID40) recalled that "I spent a lot of time reading the question, and then read the entire paragraph. And then go back to read again. First time to find a key word and the second time I read. I don't know why, but it takes time to start reading," whereas a successful reader (ID41) said "because I am sure the latter of the paragraph is not relevant about the question because last two paragraph said that obvious might be seepage dominated…It seems that *constant* and *passage* refer to something like constant water, the same meaning… because I want to find the word stream so B and C both has stream. C has a stream following so it is totally the opposite." The short response times in the sentential processing task among high scorers also suggest successful readers' effective use of expeditious reading skills.

Table 8 Eye-tracking statistics for item 12

	Total fixation duration on option A	Total fixation duration on option B	Fixation count on option A	Fixation count on option B	Total fixation duration on paragraph 6	Total fixation duration on the key area 1	Total fixation duration on the key area 2
Incorrect (*N* = 7)	1.91 (1.35)	1.86 (1.27)	6.00 (4.36)	6.29 (4.42)	9.46 (9.00)	1.11 (1.54)	.50 (1.07)
Correct (*N* = 40)	.87 (.79)	1.05 (.92)	3.08 (2.72)	3.38 (2.72)	9.59 (10/35)	1.18 (1.83)	.56 (.91)
Mann-Whitney *U*	56.00	73.00	75.00	73.00	133.00	134.00	137.50
Z	− 2.51	− 2.00	− 1.98	− 2.02	− .21	− .19	− .09
Sig. (2-tailed)	.012	.045	.048	.043	.834	.872	.932
Cohen's *d*	.79	.61	.59	.61	.06	.05	.02

Table 9 Eye-tracking statistics for item 7

	Visit count on stem	Total fixation duration on option A	Fixation count on option A	Total visit duration on option A	Visit count on option A	Attention on the first key area
Incorrect (*N* = 26)	11.04 (6.88)	5.59 (3.07)	18.04 (9.16)	6.48 (3.45)	7.5 (4.15)	.23 (.9)
Correct (*N* = 21)	7.38 (4.36)	3.22 (1.75)	11.10 (5.59)	3.72 (2.03)	4.62 (2.37)	.18 (.11)
Mann-Whitney *U*	179.50	128.00	129.00	128.50	151.50	188.00
Z	− 2.01	− 3.10	− 3.09	− 3.09	− 2.62	− 1.82
Sig. (2-tailed)	.045	.002	.002	.002	.009	.069
Cohen's *d*	.61	1.02	1.01	1.01	0.82	0.55

However, such a general tendency was not always applicable to all question types; at times, successful readers slowed down their reading rate (question 8), or unsuccessful readers were not necessarily slower than the successful readers.

The assumption that successful readers would give more attention to key phrases in text was not verified by the current data. One of possible reasons is that especially in later questions, readers likely depended on their memory rather than the text given. The interview data substantiated this impression, "Honestly I read through the paragraph when I found this not enough time so I just based on my memory I think it should be one. (ID38)" Not surprisingly, in question 6, test takers are forced to read the third paragraph for details first time, unless they skimmed it through during the first reading. Therefore, it is highly likely that test takers' eye movements directly affect their comprehension output. The amount of readers' attention on key phrases, operationalized by their eye fixation duration and visit count, does not appear to go hand in hand with the accuracy rate of comprehension. While successful readers paid more attention to the key area in question 6, it was unsuccessful readers who devoted more attention to the key area in question 7. Such discrepancy seems attributable to the question type, albeit in need of further investigation. Question 6 is the factual question in which test takers are expected to understand surface information by means of word-matching

Table 10 Eye-tracking statistics for item 9

	Fixation count on option A (key)	Fixation count on option D	Total visit duration on option A (key)	Total visit duration on option D	Total fixation duration on paragraph 5	Attention on the key sentence
Incorrect (*N* = 16)	11.50 (9.09)	11.31 (9.17)	4.25 (3.65)	4.46 (3.28)	14.15 (14.45)	.042 (.052)
Correct (N = 31)	17.35 (11.29)	6.55 (5.33)	6.40 (4.21)	2.42 (2.10)	6.39 (7.32)	.037 (.038)
Mann-Whitney *U*	159.00	151.50	161.00	133.00	166.00	234.00
Z	− 2.00	− 2.17	− 1.95	− 2.58	− 1.84	− .31
Sig. (2-tailed)	.045	.030	.051	.010	.066	.761
Cohen's *d*	.61	.67	.59	.81	.56	.09

Table 11 Eye-tracking statistics for item 1

	Total fixation duration on option B	Total fixation duration on option D (key)	Total fixation duration on paragraph 1	Fixation count on option B	Fixation count on option D (key)	Fixation count on paragraph 1	Attention on the key area
Incorrect (*N* = 17)	6.62 (3.72)	2.34 (2.25)	24.26 (11.79)	20.00 (10.33)	7.53 (7.06)	70.18 (32.69)	.19 (.11)
Correct (*N* = 30)	3.76 (4.76)	4.31 (2.56)	16.99 (9.52)	12.17 (13.77)	12.87 (8.08)	54.30 (29.67)	.22 (.19)
Mann-Whitney *U*	102.00	129.50	150.00	107.50	138.00	166.50	247.00
Z	− 3.39	− 2.78	− 2.33	− 3.27	− 2.60	− 1.96	− .18
Sig. (2-tailed)	.001	.005	.020	.001	.009	.050	.859
Cohen's *d*	1.14	.89	.72	1.08	.82	.60	.05

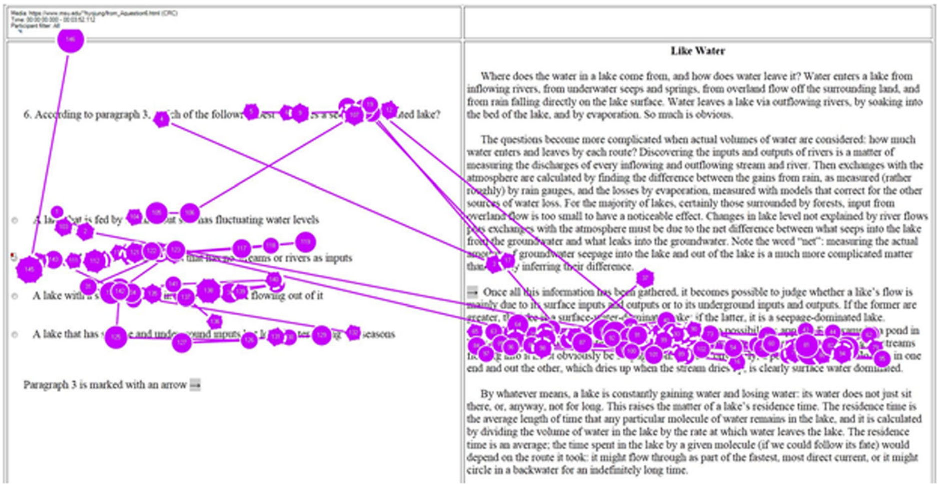
skills. In question 7, test takers are asked to make inferences, synthesizing scattered information and reading between the lines. In the latter case, longer gaze on the key areas could not necessarily contribute to the output of such high-order cognitive activities.

The current data propose the potential use of eye-tracking data for distractor analysis with sophistication. In general, unsuccessful readers likely spent more time on distractors than key options, whereas successful readers fixated their eyes on key options for a longer time. In the case of question 8, the successful readers spent more time on the key (option C) and one of the distractors (option D) than the unsuccessful readers, presumably indicating that the option D is a highly attractive distractor among high scorers. In question 12, the unsuccessful readers spent more time on the distractors A and B, meaning that for some reason, those distractors grabbed much attention among the lower scorers. Admittedly, the eye-tracking data are limited in accounting for the cognitive reasons for test takers' eye fixations.

The effect of reading comprehension subskills

Thus far, we have examined the aggregated eye movement data item by item. However, the analyses drawing on the averaged numerical data can cloud individual differences, which an eye-tracker intends to pinpoint with precision. As the relatively large standard deviations in eye-tracking measurements suggest, L2 learners seem to differ to great extent in their approaching reading questions and associated texts. The following gaze plots demonstrate considerable variation across readers, regardless of their comprehension output. Among those who got the question item 6 correct, for instance, some carefully read through the entire paragraph, while others focused only on relevant areas (see Figs. 2 and 3). Others did not even glance at the text but still found the answer correctly, presumably relying on their memory from previous questions (see Fig. 4). Among those who missed the question, some meticulously read the whole text, like some of the successful readers, while others simply failed to locate key information in text (see Figs. 5 and 6).

Learners' gaze plots themselves do not always explain how test takers derived the varying degree of comprehension output. Rather, the levels of learners' linguistic knowledge and processing skills seem to induce the differences in the graphic eye movement data. For instance, ID79, ID26, and ID82 produced relatively large CVs in the sentence



Media: https://www.msu.edu/~hyojung/from_Aquestion6.html (CRC)
Time: 00:00:00.000 - 00:03:52.112
Participant filter: All

6. Acc[illegible]ding to paragraph 3, [illegible]ch of the follow[illegible] best [illegible] s a se[illegible]ated lake?

○ A la[illegible] hat is fed by [illegible] but s[illegible] has fluctuating water levels

○ [illegible] that has no [illegible] eams or rivers as inputs

○ A lake with a s[illegible] flowing out of it

○ A lake that has s[illegible] e and unde[illegible] ound inputs [illegible] ter [illegible] seasons

Paragraph 3 is marked with an arrow →

Like Water

Where does the water in a lake come from, and how does water leave it? Water enters a lake from inflowing rivers, from underwater seeps and springs, from overland flow off the surrounding land, and from rain falling directly on the lake surface. Water leaves a lake via outflowing rivers, by soaking into the bed of the lake, and by evaporation. So much is obvious.

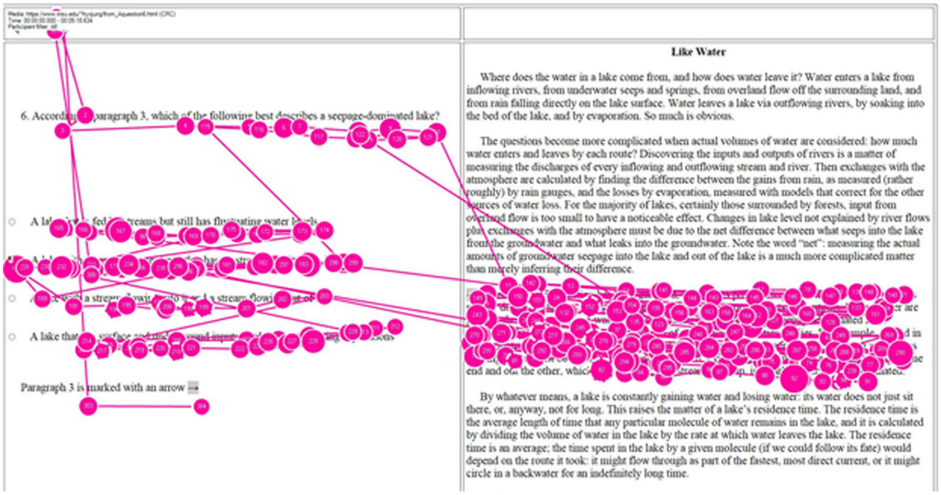
The questions become more complicated when actual volumes of water are considered: how much water enters and leaves by each route? Discovering the inputs and outputs of rivers is a matter of measuring the discharges of every inflowing and outflowing stream and river. Then exchanges with the atmosphere are calculated by finding the difference between the gains from rain, as measured (rather roughly) by rain gauges, and the losses by evaporation, measured with models that correct for the other sources of water loss. For the majority of lakes, certainly those surrounded by forests, input from overland flow is too small to have a noticeable effect. Changes in lake level not explained by river flows plus exchanges with the atmosphere must be due to the net difference between what seeps into the lake from the groundwater and what leaks into the groundwater. Note the word "net": measuring the actual amount of groundwater seepage into the lake and out of the lake is a much more complicated matter than [illegible] y inferring their difference.

→ Once all this information has been gathered, it becomes possible to judge whether a like's flow is mainly due to its surface inputs and outputs or to its underground inputs and outputs. If the former are greater, th[illegible] surface-water-dominated lake; if the latter, it is a seepage-dominated lake. [illegible] possibil[illegible] pond in [illegible] treams [illegible] obviously be [illegible] in one end and out the other, which dries up when the stream dries [illegible], is clearly surface water dominated.

By whatever means, a lake is constantly gaining water and losing water: its water does not just sit there, or, anyway, not for long. This raises the matter of a lake's residence time. The residence time is the average length of time that any particular molecule of water remains in the lake, and it is calculated by dividing the volume of water in the lake by the rate at which water leaves the lake. The residence time is an average; the time spent in the lake by a given molecule (if we could follow its fate) would depend on the route it took: it might flow through as part of the fastest, most direct current, or it might circle in a backwater for an indefinitely long time.

Fig. 3 Gaze plot from the successful candidate (ID41) on item 6

construction task, meaning less automatized syntactic processing skills; their slow sentence processing skills are realized as long fixation durations and a short distance between fixations in the gaze plot. Frequent visits and long fixation durations could also result from test takers' intention to adopt careful reading skills, matching up to the purpose of reading; however, we cannot rule out the possibility that such reading behavior is simply the corollary of learners' deficiency in L2 language ability. ID52 who demonstrated the fast processing skills at both lexical and sentential levels as well as the large knowledge of vocabulary and grammar does not seem to stay long on each word and sentence in the question, reasonably interpreted as the possession of expeditious reading skills. She did not lay her eyes on the associated paragraph but still succeeded in locating the key answer, presumably because she could have established her comprehension already during the first reading and the question 5, which is also linked to the same paragraph (Table 12).



Media: https://www.msu.edu/~hyojung/from_Aquestion6.html (CRC)
Time: 00:00:00.000 - 00:05:18.634
Participant filter: All

6. Accordi[illegible] paragraph 3, which of the following best describes a seepage-dominated lake?

○ A la[illegible] fed [illegible] reams but still has fluctuating water levels

○ [illegible]

○ [illegible] a stream [illegible] a stream flowi[illegible]

○ A lake that [illegible] surface and und[illegible]

Paragraph 3 is marked with an arrow →

Like Water

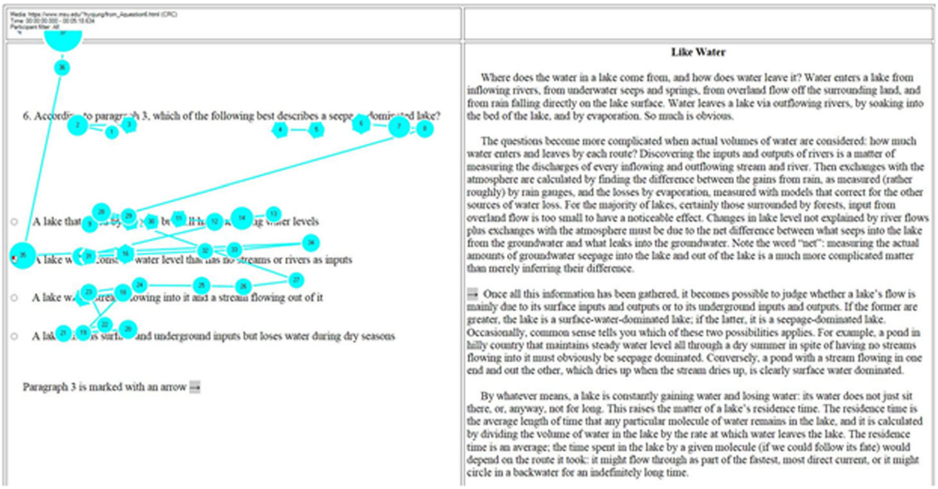
Where does the water in a lake come from, and how does water leave it? Water enters a lake from inflowing rivers, from underwater seeps and springs, from overland flow off the surrounding land, and from rain falling directly on the lake surface. Water leaves a lake via outflowing rivers, by soaking into the bed of the lake, and by evaporation. So much is obvious.

The questions become more complicated when actual volumes of water are considered: how much water enters and leaves by each route? Discovering the inputs and outputs of rivers is a matter of measuring the discharges of every inflowing and outflowing stream and river. Then exchanges with the atmosphere are calculated by finding the difference between the gains from rain, as measured (rather roughly) by rain gauges, and the losses by evaporation, measured with models that correct for the other sources of water loss. For the majority of lakes, certainly those surrounded by forests, input from overland flow is too small to have a noticeable effect. Changes in lake level not explained by river flows plus exchanges with the atmosphere must be due to the net difference between what seeps into the lake from the groundwater and what leaks into the groundwater. Note the word "net": measuring the actual amounts of groundwater seepage into the lake and out of the lake is a much more complicated matter than merely inferring their difference.

[illegible] er are [illegible] ated. [illegible] ample [illegible] d in [illegible] end and out the other, which [illegible]

By whatever means, a lake is constantly gaining water and losing water: its water does not just sit there, or, anyway, not for long. This raises the matter of a lake's residence time. The residence time is the average length of time that any particular molecule of water remains in the lake, and it is calculated by dividing the volume of water in the lake by the rate at which water leaves the lake. The residence time is an average; the time spent in the lake by a given molecule (if we could follow its fate) would depend on the route it took: it might flow through as part of the fastest, most direct current, or it might circle in a backwater for an indefinitely long time.

Fig. 2 Gaze plot from the successful candidate (ID79) on item 6



Media: https://www.msu.edu/~hyojung/from_Aquestion6.html (CRC)
Time: 00:00:00.000 - 00:05:18.634
Participant filter: All

6. According to paragraph 3, which of the following best describes a seepage-dominated lake?

○ A lake that is fed by streams but still has fluctuating water levels

● A lake with a constant water level that has no streams or rivers as inputs

○ A lake with a stream flowing into it and a stream flowing out of it

○ A lake that has surface and underground inputs but loses water during dry seasons

Paragraph 3 is marked with an arrow ➞

Like Water

Where does the water in a lake come from, and how does water leave it? Water enters a lake from inflowing rivers, from underwater seeps and springs, from overland flow off the surrounding land, and from rain falling directly on the lake surface. Water leaves a lake via outflowing rivers, by soaking into the bed of the lake, and by evaporation. So much is obvious.

The questions become more complicated when actual volumes of water are considered: how much water enters and leaves by each route? Discovering the inputs and outputs of rivers is a matter of measuring the discharges of every inflowing and outflowing stream and river. Then exchanges with the atmosphere are calculated by finding the difference between the gains from rain, as measured (rather roughly) by rain gauges, and the losses by evaporation, measured with models that correct for the other sources of water loss. For the majority of lakes, certainly those surrounded by forests, input from overland flow is too small to have a noticeable effect. Changes in lake level not explained by river flows plus exchanges with the atmosphere must be due to the net difference between what seeps into the lake from the groundwater and what leaks into the groundwater. Note the word "net": measuring the actual amounts of groundwater seepage into the lake and out of the lake is a much more complicated matter than merely inferring their difference.

➞ Once all this information has been gathered, it becomes possible to judge whether a lake's flow is mainly due to its surface inputs and outputs or to its underground inputs and outputs. If the former are greater, the lake is a surface-water-dominated lake; if the latter, it is a seepage-dominated lake. Occasionally, common sense tells you which of these two possibilities applies. For example, a pond in hilly country that maintains steady water level all through a dry summer in spite of having no streams flowing into it must obviously be seepage dominated. Conversely, a pond with a stream flowing in one end and out the other, which dries up when the stream dries up, is clearly surface water dominated.

By whatever means, a lake is constantly gaining water and losing water: its water does not just sit there, or, anyway, not for long. This raises the matter of a lake's residence time. The residence time is the average length of time that any particular molecule of water remains in the lake, and it is calculated by dividing the volume of water in the lake by the rate at which water leaves the lake. The residence time is an average; the time spent in the lake by a given molecule (if we could follow its fate) would depend on the route it took: it might flow through as part of the fastest, most direct current, or it might circle in a backwater for an indefinitely long time.

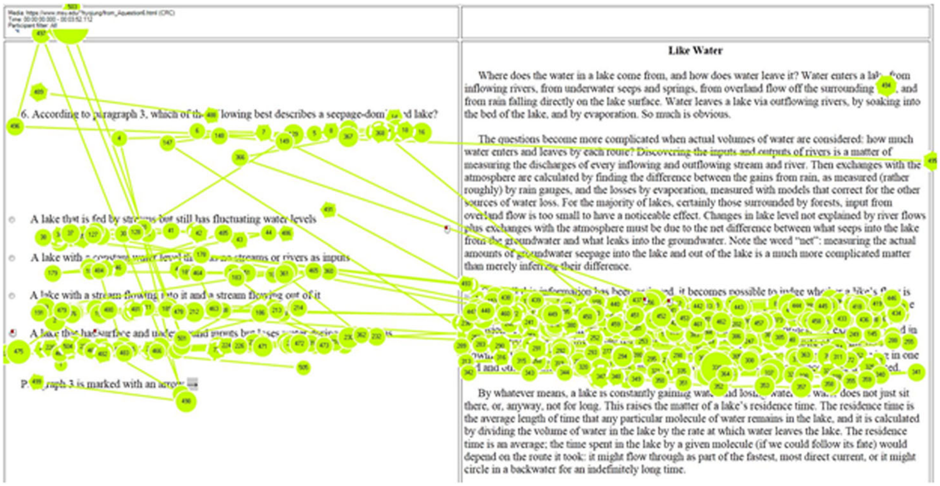
Fig. 4 Gaze plot from the successful candidate (ID52) on item 6

Discussion

What does the iBT TOEFL reading section measure?

Assuming that reading is a purposeful behavior, the iBT TOEFL claims to measure academic reading skills, including *reading to find information*, *reading for basic comprehension*, and *reading to learn* (ETS, 2009). From a process-based perspective, *reading to find information* is to involve expeditious reading skills at both local and global level (e.g., scanning, searching, and skimming), whereas *reading for basic comprehension* requires careful reading skills for both explicit and implicit information. The current study aimed to investigate to what extent the TOEFL reading test successfully elicited the intended reading processes.

Overall, high scorers tended to read fast compared to low scorers; such disparity seems better observed in factual questions than in inference questions. Learners' eye movements, especially fixation duration and visit counts, may effectively account for



Media: https://www.msu.edu/~hyojung/from_Aquestion6.html (CRC)
Time: 00:00:00.000 - 00:03:52.112
Participant filter: All

6. According to paragraph 3, which of the following best describes a seepage-dominated lake?

○ A lake that is fed by streams but still has fluctuating water levels

○ A lake with a constant water level that has no streams or rivers as inputs

○ A lake with a stream flowing into it and a stream flowing out of it

● A lake that has surface and underground inputs but loses water during dry seasons

Paragraph 3 is marked with an arrow ➞

Like Water

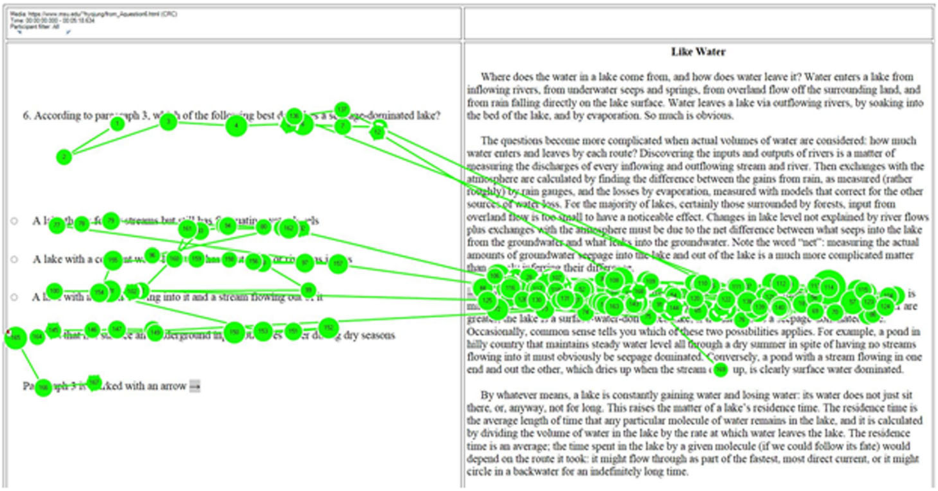
Where does the water in a lake come from, and how does water leave it? Water enters a lake from inflowing rivers, from underwater seeps and springs, from overland flow off the surrounding land, and from rain falling directly on the lake surface. Water leaves a lake via outflowing rivers, by soaking into the bed of the lake, and by evaporation. So much is obvious.

The questions become more complicated when actual volumes of water are considered: how much water enters and leaves by each route? Discovering the inputs and outputs of rivers is a matter of measuring the discharges of every inflowing and outflowing stream and river. Then exchanges with the atmosphere are calculated by finding the difference between the gains from rain, as measured (rather roughly) by rain gauges, and the losses by evaporation, measured with models that correct for the other sources of water loss. For the majority of lakes, certainly those surrounded by forests, input from overland flow is too small to have a noticeable effect. Changes in lake level not explained by river flows plus exchanges with the atmosphere must be due to the net difference between what seeps into the lake from the groundwater and what leaks into the groundwater. Note the word "net": measuring the actual amounts of groundwater seepage into the lake and out of the lake is a much more complicated matter than merely inferring their difference.

➞ Once all this information has been gathered, it becomes possible to judge whether a lake's flow is mainly due to its surface inputs and outputs or to its underground inputs and outputs. If the former are greater, the lake is a surface-water-dominated lake; if the latter, it is a seepage-dominated lake. Occasionally, common sense tells you which of these two possibilities applies. For example, a pond in hilly country that maintains steady water level all through a dry summer in spite of having no streams flowing into it must obviously be seepage dominated. Conversely, a pond with a stream flowing in one end and out the other, which dries up when the stream dries up, is clearly surface water dominated.

By whatever means, a lake is constantly gaining water and losing water: its water does not just sit there, or, anyway, not for long. This raises the matter of a lake's residence time. The residence time is the average length of time that any particular molecule of water remains in the lake, and it is calculated by dividing the volume of water in the lake by the rate at which water leaves the lake. The residence time is an average; the time spent in the lake by a given molecule (if we could follow its fate) would depend on the route it took: it might flow through as part of the fastest, most direct current, or it might circle in a backwater for an indefinitely long time.

Fig. 5 Gaze plot from the unsuccessful candidate (ID26) on item 6



6. According to para[illegible]ph 3, wh[illegible] of the foll[illegible]ing best d[illegible] s a se[illegible]age-dominated lake?

○ A l[illegible] th[illegible] fe[illegible] streams but still has f[illegible]ating [illegible] els

○ A lake with a c[illegible] at w[illegible] has [illegible] or [illegible]

○ A l[illegible] with a[illegible] ing into it and a stream flowing ou[illegible] it

[illegible] that [illegible] su[illegible]e an [illegible]derground in[illegible] es [illegible] er d[illegible] g dry seasons

Pa[illegible] ph 3 is [illegible]ked with an arrow ➡

Like Water

Where does the water in a lake come from, and how does water leave it? Water enters a lake from inflowing rivers, from underwater seeps and springs, from overland flow off the surrounding land, and from rain falling directly on the lake surface. Water leaves a lake via outflowing rivers, by soaking into the bed of the lake, and by evaporation. So much is obvious.

The questions become more complicated when actual volumes of water are considered: how much water enters and leaves by each route? Discovering the inputs and outputs of rivers is a matter of measuring the discharges of every inflowing and outflowing stream and river. Then exchanges with the atmosphere are calculated by finding the difference between the gains from rain, as measured (rather roughly) by rain gauges, and the losses by evaporation, measured with models that correct for the other sources of water loss. For the majority of lakes, certainly those surrounded by forests, input from overland flow is too small to have a noticeable effect. Changes in lake level not explained by river flows plus exchanges with the atmosphere must be due to the net difference between what seeps into the lake from the groundwater and what leaks into the groundwater. Note the word "net": measuring the actual amounts of groundwater seepage into the lake and out of the lake is a much more complicated matter than [illegible] inferring their difference.

[illegible] is [illegible] are greate[illegible] the lake [illegible] a surfa[illegible]-water-dom[illegible] a seepage-[illegible]. Occasionally, common sense tells you which of these two possibilities applies. For example, a pond in hilly country that maintains steady water level all through a dry summer in spite of having no streams flowing into it must obviously be seepage dominated. Conversely, a pond with a stream flowing in one end and out the other, which dries up when the stream [illegible] up, is clearly surface water dominated.

By whatever means, a lake is constantly gaining water and losing water: its water does not just sit there, or, anyway, not for long. This raises the matter of a lake's residence time. The residence time is the average length of time that any particular molecule of water remains in the lake, and it is calculated by dividing the volume of water in the lake by the rate at which water leaves the lake. The residence time is an average; the time spent in the lake by a given molecule (if we could follow its fate) would depend on the route it took: it might flow through as part of the fastest, most direct current, or it might circle in a backwater for an indefinitely long time.

Fig. 6 Gaze plot from the unsuccessful candidate (ID82) on item 6

their comprehension of surface information; without looking at relevant areas, test takers may well fail to identify facts explicitly stated in the text. Little or no relation between learners' reading rate and their comprehension for implicit information could provide evidence in support of the cognitive validity of inference questions; as fast readers likely slow down for higher-order thinking, differences between successful and unsuccessful readers could become obscured in respect of reading speed. However, the jury is still out on whether the TOEFL questions in effect provoked the expeditious reading skills. In some factual questions, successful readers were not always faster in processing a reading passage. In the first reading page, presumably the only place where learners can exert their expeditious reading skills across paragraphs, both high and low scorers did not read the text properly, rushing into the first question on the subsequent page. The actual TOEFL iBT test does not allow test takers to move to the first question without scrolling down to the bottom of the text, which was not programmed in the present study. Apart from the technical issue, there seems to be a clear tendency for test takers to opt out the first reading page; only 30% of high scorers skimmed through the entire text. Such behaviors could be the manifestation of learners' use of test-taking strategies in an attempt to save time for challenging questions. As results, careful reading at the local level seems to prevail in the TOEFL reading test, primarily verified by the reading time, the saccade length between fixation points, and the fixation count. Test takers are asked to read each paragraph (or less than a paragraph) carefully in orderly fashion, taking a comprehension question as guidance to process

Table 12 The sub-reading abilities of each test taker from item 6

	Item 6	Vocabulary knowledge	Grammar knowledge	Lexical processing skills	Sentential processing skills	Working memory	Total reading scores
ID79	Correct	67/90	26/35	.45	.50	23	11/14
ID41	Correct	74/90	27/35	.31	.47	18	11/14
ID52	Correct	88/90	33/35	.18	.37	34	14/14
ID26	Incorrect	72/90	16/35	.24	.52	27	6/14
ID82	Incorrect	51/90	17/35	.34	.59	15	6/14

the text. In the final summary question, which ostensibly calls for global reading, many test takers seem to rely on their memory; none of the high scorers glanced at the reading passage to solve the question. In Cohen and Upton (2007), the international students reported that the reading-to-learn questions were relatively easy. The authors reasoned that test takers are to become too familiar with the reading passage by the time they reached the last question. Owing to the repeated readings induced by previous questions, test takers could feel that they already know the key ideas too well to take a second look at the text. Otherwise, they might have been simply running out of time for the last question.

Assessing reading fluency is crucial especially for advanced learners. Fluency helps to reduce learners' cognitive load for linguistic processing at every level, thus enabling readers to focus more on meaning (Logan, 1997). Empirical studies have reported a significant correlation between reading speed and reading comprehension (e.g., Fuchs et al., 2001). Good L2 readers have been found to have faster word recognition and automatic syntactic processing skills than poor L2 readers (Nassaji 2003; Grabe and Stoller 2002). Note that reading fluency entails both accuracy and automaticity; without accurate comprehension, reading speed alone cannot account for fluent reading skills. According to the ACTFL (American Council on the Teaching of Foreign Languages) proficiency guidelines, fluency becomes the determining factor as learners advance to higher levels; superior learners are described as being able to read diverse texts *efficiently*. Because fluency is acquired in the latter stage of L2 development (e.g., Anderson's ACT-R theory), it would be appropriate to administer fluency measures to advanced learners, such as those preparing to enter US universities. Language testing researchers have also voiced the need of assessing reading fluency; Alderson (2000) criticized the current comprehension test for not taking speed into account and backs timed readings as a means of diagnosing learners' development of automaticity. Weir et al. (2009) also indicated international students' struggle with fast reading in the academic setting. Taken together, because (a) fluency is an important constituent of the reading construct, (b) academic reading requires rapid reading skills, and (c) the target population of such high-stake tests are usually advanced learners, it seems high time for high-stakes, academic tests to incorporate or enhance testing reading fluency.

The other point of discussion is the purpose of vocabulary question items. The TOEFL vocabulary question items are supposed to measure learners' inferencing ability, i.e., guessing meaning from context. Of three vocabulary questions, only one turned out to be valid as a result of item analysis. The eye-tracking analyses also challenged the cognitive validity of vocabulary questions; nearly one third (30%) of the participants were able to find the correct answers without looking at the text. Even in the case where learners fixated their eyes on the text, reading was often limited to a phrase or a word level, presumably indicating that test takers may turn to their knowledge of collocation by checking the words adjacent to a target word. Moreover, little or no group difference was found in terms of test takers' reading time on the associated paragraph and the amount of attention on key phrases. In Cohen and Upton (2007), test takers, most of whom were from East Asian countries, reported that they utilized their vocabulary knowledge if they already knew the target word. Similar ideas were echoed in the current stimulated interview; ID38 said, "When I see the word *considerable*, because I know this word, so I know I just check which one is the meaning, like, *considerable* so once I found, I think I choose A." ID51 also noted "Because it's *further* and the only option is *additional*, I didn't look at… I found out when I read the TOEFL test, it may be a bad habit. It's just like fifty-fifty chance sometimes you don't have to look at the passage you can just choose answer but sometimes it didn't work." Overall, vocabulary questions may have succeeded in differentiating between high and low scorers, in that good readers are likely to have larger vocabulary size (Jeon & Yamashita, 2014). However, whether the question items effectively tapped into the intended reading skills—guessing meaning from context—remains questionable.

What can the eye-tracking data reveal about cognitive validity?

The value of eye-tracking data, triangulated with other qualitative research methods, primarily lies in uncovering test takers' subconscious reading processes, including the direction and the amount of their attention. However, in the context of L2 testing, longer fixation duration does not necessarily represent more amount of learners' attention; it could presumably result from learners' deficiency in linguistic knowledge, inefficiency in processing skills, or the varying degree of complexity of language itself (Rayner, 2009). Not surprisingly, in quite a few question items, there was little or no significant group difference in terms of learners' eye movements, both qualitatively and quantitatively (Bax & Chan, 2019; McCray & Brunfaut, 2018). In a similar vein, the collected data in the current study also failed to meet the fundamental assumption that successful readers would be efficient in reading and locating key information and better attend to key phrases; on some items, successful readers were slower than unsuccessful readers in reading and did not necessarily pay more attention to key information. In Bax (2013), only five out of 11 items succeeded in showing group differences in terms of test takers' reading processes. The author reasoned that successful candidates might not necessarily go through more cognitive processes different from unsuccessful ones, or that they might employ some other strategies that cannot be traced by eye-tracking technology. That said, we should not be misled into believing that the items that did not yield a statistical group difference are cognitively invalid. L2 readers have different profiles of strengths and weaknesses in comprehension subskills. As far as such individual differences are concerned, not all successful readers may well take the same trajectory to obtain the same outcome of comprehension (see Figs. 2, 3, and 4). Likewise, reasons can vary why unsuccessful readers missed a question, some of which may not be reflected in their eye movements.

Bax (2013) highlighted that the cognitively valid items, identified by eye movement analyses, involved only lexical or grammatical processing. This is reminiscent of the fact that a number of reading studies have limited the use of an eye-tracker to investigating learners' word recognition or sentential processing skills, although some have tried to expand the application of eye-tracking study beyond the sentential level (e.g., Rayner et al., 2006). The gaze plots illustrated earlier also suggest that learners' eye-tracking data are susceptible to their bottom-up processing skills, regardless of their comprehension outcome; the successful reader in Fig. 2 and the unsuccessful reader in Fig. 5, both of whom were slow in the sentence processing task, exhibited the similar pattern of eye movements, including frequent visit counts and short saccade lengths between fixations. In a similar vein, the items that made relatively distinct group differences in eye movements were often limited to factual questions, which primarily demand locating and matching skills at the lexical level. Admittedly, we could not observe any meaningful group differences in text reading associated with inference questions, which involve more than L2 language processing skills.

The application of eye-tracking technology to language testing has emerged in recent years. Thus, the method of using eye movement data is still explorative for the purpose of test validation. Caution needs to be taken in interpreting the results and relating them to validity arguments. It is also noteworthy that test validity is a matter of degree (Brown, 2005); thus, it would be unreasonable to appraise the validity of the test in a dichotomous manner. For some of the items, the eye movement data successfully differentiated good from poor readers, which can effectively defend the validity of the test items. In other items, unexpected patterns turned up, which may question the validity of the item, but this difference is not sufficient to undermine the entire test's validity. Additional evidence, such as from the stimulated recall data, would be necessary to investigate the reasons for test takers' peculiar eye movements. Lastly, the present study suggests the potential use of eye-tracking analysis for the purpose of distractor analysis. At present, distractor analysis is based on the final responses that test takers mark on the answer sheet, which may veil their decision-making process. As test takers' attraction to a distractor does not necessarily lead to their final decision, eye movement data would help evaluate the quality of the response choices more accurately.

Conclusion

The present study set out to explore the validity evidence for the iBT TOEFL reading test by observing the cognitive processes involved in the test. The eye-recording analysis has led to the following general observations; test takers tend to rely heavily on the careful reading skills, whereas the expeditious reading skills were rarely activated. The scope of reading was often restricted to the local level; learners hardly read more than a paragraph to answer questions. In some factual question items, successful readers were more efficient in reading and locating key information, whereas we could not notice such group differences in other items. Lastly, the gaze plots hinted that learners' eye movements manifest the various form of interactions among comprehension subskills, chiefly subject to bottom-up linguistic processing. Interpretations can vary in light of the judgment of cognitive validity, however. Such findings can challenge the cognitive validity of the test, as it fails to fully satisfy *similarity* and *comprehensiveness*, the important characteristics of cognitive validity proposed by Field (2013). Others who view L2 reading as a language problem (e.g., Jeon & Yamashita, 2014), meaning that the knowledge of L2 vocabulary and grammar has the primary responsibility for L2 reading comprehension, could admit that the test is valid enough to measure the essence of learners' reading ability. Note that high scorers in this study showed a larger knowledge of vocabulary and grammar than low scorers.

Findings yield practical implications for test development. As an increasingly number of scholars, teachers, and students acknowledge the importance of L2 reading fluency in a post-secondary education, test developers and item writers need to devise effective ways of assessing learners' fast reading skills beyond the paragraph level. Rearranging item types can make a quick and simple change to test takers' reading patterns; placing skimming items in front may adequately induce learners to deploy expeditious reading skills, while minimizing the involvement of other reading strategies and skills. Otherwise, using hyperlinks, which allow reading across texts in a test setting, could be a technical solution. With regard to a vocabulary item, its item specification needs to be reconsidered. Evidence seems scant that test takers harnessed their inferencing skills to solve vocabulary items. By carefully selecting a test word and a context where the word is situated, we can prevent the inferencing question item from assessing one's vocabulary size.

There are a number of limitations inherent in the experiment design and data analysis. Participants did not take the full set of the TOEFL iBT reading test, primarily due to time constraints. As the reading passages used for the experiment are limited to the expository texts, learners' reading behaviors observed in the eye movements might not be generalizable to the test associated with the historical or argumentative text. There is an ecological issue as well; albeit with an attempt to closely replicate the real testing setting, the one-to-one lab-based data collection procedure might have to some extent affected learners' test-taking behaviors. Also, the participants could be too homogeneous to reveal group differences; except for a couple of provisional students, they were all regular international students who met the minimum requirement of the standardized English test. Finally, the use of Times New Roman could have been a source of errors in analyzing the eye-tracking data, as it is not a fixed-width font.

Despite the limitations, the present study still contributes to enriching validity arguments by providing additional and concurrent empirical evidence, while suggesting a great potential of the eye-tracking analysis for test validation. Future research should further delve into the analysis methods, presumably in quantifying the qualitative nature of eye movement data for group comparisons, expanding the application of eye-tracking technology to beyond the lexical level, and relating other qualitative data (e.g., interviews and questionnaires) to eye-tracking analysis. It is hoped that the present study, far from complete, can serve as a stepping stone for future work in the direction of looking into test takers' cognition and thus fleshing out validity arguments.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s40468-020-00107-0.

Additional file 1. Key information in the reading passage.

Acknowledgements
I would like to thank anonymous reviewers for their insightful comments on an earlier draft of this paper.

Author's contributions
HL is the sole contributor to this research paper. The author read and approved the final manuscript.

Funding
The study was supported by a grant (DDG 2012) from the International Research Foundation for English Language Education.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests
The author declares that she has no competing interests.

Received: 3 May 2020 Accepted: 17 August 2020
Published online: 13 October 2020

References

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

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

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

Bax, S., & Chan, S. (2019). Using eye-tracking research to investigate language test validity and design. *System*.

Bax, S., & Weir, C. (2012). Investigating learners' cognitive reading processes during a computer-based CAE reading test. *University of Cambridge ESOL Examinations Research Note*, *47*, 3–14.

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice*, *17*, 37–45. https://doi.org/10.1111/j.1745-3992.1998.tb00627.x.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment (New ed.)*. New York: McGraw-Hill.

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

Chan, S.H.C. (2013) *Establishing the validity of reading-into-writing test tasks for the UK academic context*. PhD dissertation. University of Bedfordshire.

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

Cohen, A. D., & Upton, T. A. (2007). `I want to go back to the text': Response strategies on the reading subtest of the new TOEFL(R). *Language Testing*, *24*(2), 209–250. https://doi.org/10.1177/0265532207076364.

Conklin, K., Alotaibi, S., Pellicer-Sánchez, A., & Vilkaitė-Lozdienė, L. (2020). What eye-tracking tells us about reading-only and reading-while-listening in a first and second language. *Second Language Research*, 0267658320921496.

Educational Testing Service (2009). *The official guide to the TOEFL test*, (3rd ed.,). New York: McGraw-Hill.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of experimental psychology: General*, *128*(3), 309–331.

Enright, M. K., & Schedl, M. (2000). *Reading for a reason: Using reader purpose to guide test design. Unpublished manuscript*. Princeton: Educational Testing Service.

Field, J. (2011). Into the mind of the academic listener. *Journal of English for Specific Purposes*, *10*, 102–112.

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

Field, J. (2013). Cognitive validity. *Examining Listening: Research and Practice in Assessing Second Language Listening, Studies in Language Testing*, *35*, 77–151.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific studies of reading*, *5*(3), 239–256.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist*, *36*, 923–936.

Godfroid, A. (2019). Investigating instructed second language acquisition using L2 learners' eye-tracking data. In R. P. Leow (Ed.), *The Routledge Handbook of Second Language Research in Classroom Learning*, (pp. 44–57). New York: Routledge.

Godfroid, A. (2020). *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide*. New York: Routledge.

Godfroid, A., & Hui, B. (2020). Five common pitfalls in eye-tracking research. *Second Language Research*, *36*(3), 277–305 https://doi.org/10.1177/0267658320921218.

Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educational Researcher*, *36*(8), 456–462.

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

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

Henning, G. (1987). *A Guide to Language Testing*. Cambridge: Newbury House.

Jamieson, J. M., Eignor, D., Grabe, W., & Kunnan, A. J. (2008). Frameworks for a new TOEFL In. In C. Chapelle, M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language*, (pp. 55–95). New York: Routledge.

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*(1), 160–212. https://doi.org/10.1111/lang.12034.

Judith, R. (1995). Student responses to reading strategies instruction. *Foreign Language Annals*, *28*(2), 262–273.

Juhasz, B. J., & Rayner, K. (2006). The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition*, *13*(7-8), 846–863.

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

Khalifa, H., & Weir, C. J. (2009). *Examining Reading: Research and Practice in Assessing Second Language Reading*. Cambridge University Press.

Killen, R. (2003). Validity in outcomes-based assessment. *Perspectives in Education*, *21*(1), 1–14.

Laubrock, J., & Kliegl, R. (2015). The eye-voice span during reading aloud. *Frontiers in psychology*, *6*, 1432–1451.

Lemaire, B., Guerin-Dugue, A., Baccino, T., Chanceaux, M., & Pasqualotti, L. (2011). A cognitive computational model of eye movements investigating visual strategies on textual material. In *33rd annual meeting of the Cognitive Science Soceity*, (pp. 1146–1151). Boston: Massachusetts.

Lim, H., & Godfroid, A. (2015). Automatization in second language sentence processing: A partial, conceptual replication of Hulstijn, Van Gelderen, and Schoonen's 2009 study. *Applied Psycholinguistics*, *36*(5), 1247–1282.

Logan, G. D. (1997). Automaticity and reading: Perspectives from the instance theory of automatization. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, *13*(2), 123–146.

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

Morrison, R. E. (1984). Manipulation of stimulus onset delay in reading: evidence for parallel programming of saccades. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(5), 667–682.

Nassaji, H. (2003). Higher–level and lower–level text processing skills in advanced ESL reading comprehension. *The Modern Language Journal*, *87*(2), 261–276.

Overmann, M. (1999). Reading is guided creation. *Fremdsprachenunterricht*, *43*(5), 327–332.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing*, *20*(1), 25–56.

Pollatsek, A., Reichle, E. D., & Rayner, K. (2006). Tests of the E-Z Reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology*, *52*, 1–56.

Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability test takers: a structural equation modelling approach. *Language Testing*, *15*(3), 333–379. https://doi.org/10.1177/026553229801500303.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly journal of experimental psychology*, *62*(8), 457–506. https://doi.org/10.1080/17470210902816461.

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

Rayner, K., Pollatsek, A., Ashby, J., & Clifton Jr., C. (2012). *Psychology of reading*. Psychology Press.

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment*, *28*(3), 164–171. https://doi.org/10.1027/1015-5759/a000123.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*(1), 125.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, *26*(1), 5–30.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the Vocabulary Levels Test. *Language testing*, *18*(1), 55–88.

Segalowitz, N., & de Almeida, R. G. (2002). Conceptual representation of verbs in bilinguals: semantic field effects and a second-language performance paradox. *Brain and Language*, *81*(1-3), 517–531. https://doi.org/10.1006/brln.2001.2544.

Segalowitz, N., & Frenkiel-Fishman, S. (2005). Attention control and ability level in a complex cognitive skill: attention shifting and second-language proficiency. *Memory & Cognition*, *33*(4), 644–653 Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16248329.

Sereno, S. C., & Rayner, K. (2003). Measuring word recognition in reading: eye movements and event-related potentials. *Trends in cognitive sciences*, *7*(11), 489–493.

Shiotsu, T. (2003). *Linguistic knowledge and processing efficency as predictors of L2 reading ability: a component skills analysis*. Unpublished PhD thesis, The University of Reading.

Song, X., & Cheng, L. (2009). Language learner strategy use and test performance of Chinese learners of English. *Language Assessment Quarterly*, *3*(3), 37–41.

Stricker, L. J., & Rock, D. A. (2008). Factor Structure of the TOEFL Internet-Based Test across Subgroups. *ETS Research Report Series*, *2008*(2), i-38.

Thelk, A. D., & Hoole, E. R. (2006). What are you thinking? Postsecondary student think-alouds of scientific and quantitative reasoning items. *The Journal of General Education*, *55*(1), 17–39.

Verhoeven, L. T. (2000). Components in early second language reading and spelling. *Scientific Studies of Reading*, *4*, 313–330.

Weir, C., Hawkey, R., Green, A., Unaldi, A., & Devi, S. (2009). The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university. *IELTS Research Reports*, *9*, 97–156.

Weir, C. J. (2005). *Language testing and validation*. Hampshire: Palgrave McMillan.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.