International Journal of TESOL Studies (2022)
Vol. 4 (1) 83-104 https://doi.org/10.46451/ijts.2022.01.07

Article

Eye-Tracking L2 Students Taking Online Multiple-Choice Reading Tests: Benefits and Challenges

Nicola Latimer*
University of Bedfordshire, UK

Sathena Chan
University of Bedfordshire, UK

Received: 15 September 2021/Accepted: 31 January 2022/Published: 30 March 2022

Abstract
Recently, there has been a marked increase in language testing research involving eye-tracking. It appears to offer a useful methodology for examining cognitive validity in language tests, i.e., the extent to which the mental processes that a language test elicits from test takers resemble those that they would employ in the target language use domains. This article reports on a recent study which examined reading processes of test takers at different proficiency levels on a reading proficiency test. Using a mixed-methods approach, the study collected cognitive validity evidence through eye-tracking and stimulated recall interviews. The study investigated whether there are differences in reading behaviour among test takers at CEFR B1, B2 and C1 levels on an online reading task. The main findings are reported and the implications of the findings are discussed to reflect on some fundamental questions regarding the use of eye-tracking in language testing research.

Keywords
Eye-tracking, reading tests, cognitive validity, multiple-choice

1 Introduction

A thorough understanding of the cognitive processes used by L2 students engaged in reading test tasks offers an opportunity to gather evidence to support the validity argument of the tests as well as to inform learning oriented assessment. Recently, there has been a marked increase in language testing research involving eye-tracking. It appears to offer a useful methodology for examining cognitive validity in language tests, i.e., the extent to which the mental processes that a language test elicits from test takers resemble those that they would employ in the target language use domains (Weir, 2005). This article reports on a recent study which examined reading processes of test takers at different proficiency levels on a reading proficiency test. Using a mixed-methods approach, the study collected cognitive validity evidence through eye-tracking and stimulated recall interviews. The study investigated whether there

*Corresponding author. Email: nicola.latimer@beds.ac.uk

are differences in reading behaviour among test takers at CEFR B1, B2 and C1 levels on a reading task. The main findings are reported and the implications of the findings are discussed to reflect on some fundamental questions regarding the use of eye-tracking in language testing research.

2 Literature Review

2.1 Cognitive validity

Weir's (2005) socio-cognitive framework advocates the importance of cognitive validity in relation to other validity components when establishing validity argument for language tests. The cognitive validity of a language test concerns the extent to which a test elicits the same cognitive processes as those likely to be demanded of the test takers in the target language use (TLU) domain (Bachman & Palmer, 1996) beyond the test (Glaser, 1991). The concept of cognitive validity, recognised by validation frameworks such as 'Standards for educational and psychological testing' (AERA, APA, NCME, 2014, p. 15) can be traced back to Messick (1989) and Embretson (1983). While cognitive validity is a key aspect of test validity, the number of studies dedicated to investigating test takers' processes on tasks remains small (Weir, Vidakovic & Galaczi, 2013). Therefore, language test providers need to demonstrate evidence showing which cognitive processes their tasks elicit and profiling how test takers at different levels actually use these processes to complete the tasks. The socio-cognitive framework also includes context validity which concerns the contextual parameters of a task. These parameters are often socially or externally determined in terms of the demands of the task setting, with its specified input and expected output. To establish which cognitive processes are involved in reading, a review of the literature relating to models of reading was carried out.

2.2 Models of reading

Khalifa and Weir's (2009) model of reading is regarded as suitable for the purpose of test validation studies as the model draws on external evidence from cognitive psychology concerning the nature of reading that examining boards should aim to sample through test tasks. The principal concern is with the mental processes readers actually use in comprehending texts when engaging in different types of real-life reading.

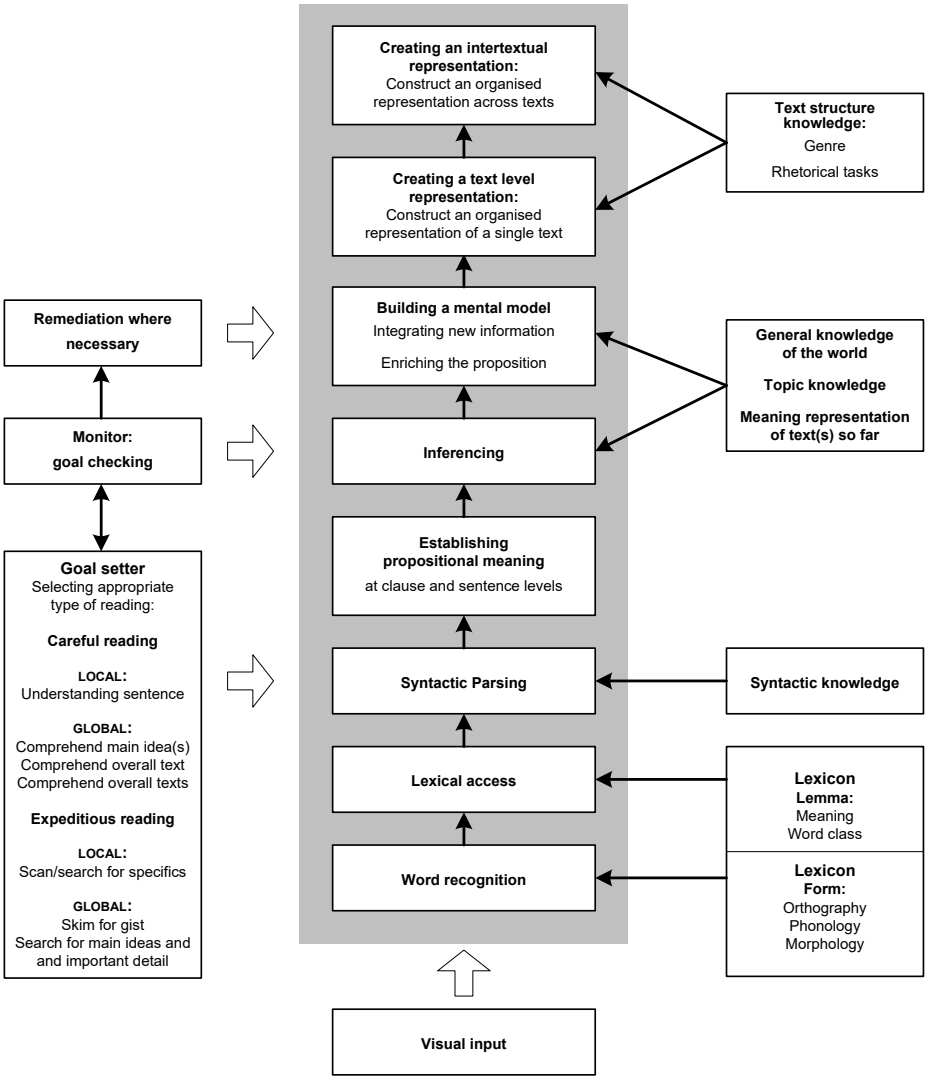
The various types of reading and the cognitive processes language tasks may give rise to are represented in Figure 1. The left hand column represents the metacognitive activity of goal setting. The goal setter is responsible for deciding what *type(s) of reading*, e.g., careful expeditious reading, to employ when faced with a text or texts in order to achieve the reader's goals.

Careful reading is intended to extract meaning from presented material at a local or a global level, i.e., within or beyond the sentence right up to the level of the complete text or texts. There is also a case for taking account of the speed of reading as well as comprehension. Studies into students' reading abilities have indicated that for many readers, reading quickly, selectively and efficiently (i.e., expeditious reading) poses greater problems than reading carefully and efficiently (Beard, 1972; Weir, 1983; Weir, Yang & Jin, 2000). Expeditious reading of continuous prose to access desired information in a text is difficult for many L2 readers because it demands rapid recognition which is contingent upon sufficient practice in reading in the target language. Expeditious reading includes skimming, search reading, and scanning.

Critical decisions taken during goal setting in turn affect the *level(s) of processing* to be activated in the central column of the model. The three processes at the bottom: word recognition (decoding), lexical access, syntactic parsing enable a reader to build a surface representation of text, i.e. decoding words and recognising syntactic patterns of these words.

Figure 1

Khalifa and Weir's Model of Reading (2009:43)



The process of establishing propositional meaning enables a reader to build a representation of the meaning of clauses and sentences. This representation is still largely text-based, i.e. meaning closely tied to the text's formulation. These lower-level processes would allow a reader to recognise the linguistic properties of the text and the literal propositions (meaning) of these clauses and sentences. Nevertheless, for meaningful comprehension to take place, a reader needs to interact with the text by supplying information not explicitly expressed by the writer at times (inferencing), and by means of world knowledge and awareness of the current topic (building a mental situation model). Creating a text-level representation and creating an intertextual representation are the highest-level reading process which a reader uses to work out conceptual links between propositions and build a coherent understanding of the text as a whole or across multiple texts. *Monitor* can be applied to each level of processing that is activated in response to the goal setter's instructions.

The knowledge base required for comprehension constitutes the right hand column. The knowledge based cognitive load of the text to be processed is largely determined by task features (related to contextual validity of the test). Grabe and Jiang (2014) identify a list of 12 factors which impact L2 learners' comprehension ability. Similar to Khalifa & Weir's (2009) model of reading, Grabe and Jiang's inventory recognises the multiple-dimensional nature of L2 reading comprehension. Their list of L2 reading comprehension ability can also be broadly interpreted in terms of reading sub-skills (e.g.

phonological, orthographical, parsing), knowledge base (e.g. vocabulary knowledge) and cognitive processes (e.g. goal-setting, inferencing).

Bernhardt (2005) suggests that adult L2 readers, who are skilled in L1 reading, may use strategies and skills from their L1 to help compensate for a lack of automation in lower level L2 reading skills. This tactic can achieve accurate compensation, but more slowly than skilled L1 readers. This approach echoes that of Walczyk (2000) who suggests that slow but accurate L1 reading could also be accounted for by the use of higher level processes to supplement or subsidise a lack of automation in low level processes.

As L2 reading comprehension ability involves a range of sub-skills, knowledge base and cognitive processes, Alderson (2000) argues that that test developers need to be careful in considering lists or taxonomies of skills. Test developers need to reflect on the tasks in light of process, product and test taker. In this article, we report a study which utilised eye tracking and stimulated recall to map the evidence of reading behaviours onto Khalifa and Weir's (2009) model of L2 reading in relation to goal setter (left hand column in Fig. 1) and level of cognitive processing (central column in Fig. 1).

Having considered the concept of cognitive validity and models of reading, we move on to review the literature relating to eye-tracking and the use of eye-tracking for test validation purposes.

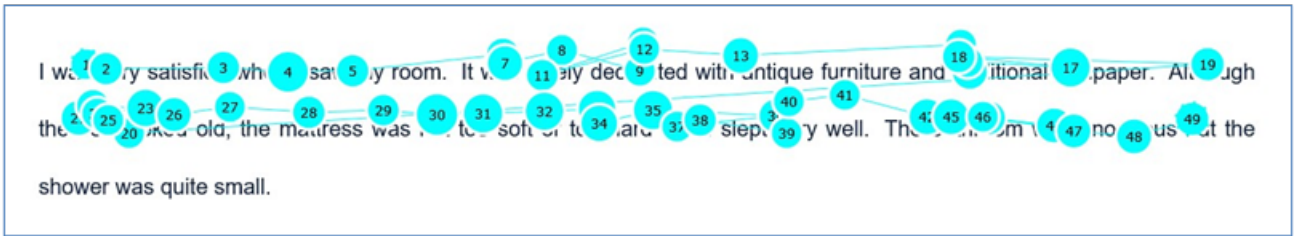
2.3 The work of the eyes during reading

Eye-tracking has revealed the fine detail of how our eyes move during reading. When people read, whilst it may seem that the eyes glide smoothly along the line of text that is not the case. In fact, the eyes make a series of jumps along the line interspersed with pauses (Holmqvist, Nyström, Andersson, Dewhurst, Jarodzka & Van de Weijer.,2011; Rayner, Pollatsek, Ashby & Clifton, 2012). The jumps are called saccades and the pauses are called fixations. These two measures of eye movements during reading are central to much of the research which has been conducted using eye-tracking.

The characteristics of fixations and saccades have been established through extensive research some of which dates back to the late 1800s when the first mechanical eye-trackers were developed. During fixations, which typically last about 250 milliseconds, although the range can vary from just 66 milliseconds to 416 milliseconds (Rayner et al., 2012, p. 93), the eye remains relatively still. The eye is not absolutely still but rather trembling, making very tiny adjustments (nystagmus) to keep the focus of the eye at the same location on the page / screen. The eye then moves extremely quickly during the saccade to the next fixation. Typically, saccades last just 40 milliseconds (ibid). Wolverton and Zola (1983) demonstrated that during saccades, no visual information is registered; the eye is moving too quickly. On average, reading saccades move forward 8-character spaces through the text (Holmqvist et al., 2011; Rayner et al., 2012). Not every word is fixated, with high frequency words and predictable words more likely to be skipped (Blanchard, Pollatsek, & Rayner; 1989; Brysbaert & Vitu, 1989).

Figure 2

Diagram of Reading Fixations



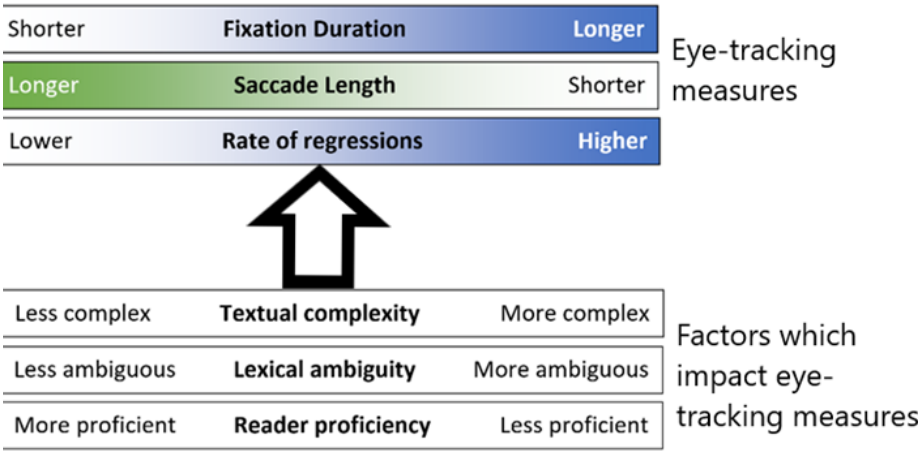
I was very satisfied when I saw my room. It was nicely decorated with antique furniture and traditional wallpaper. Although the bed looked old, the mattress was not too soft or too hard and I slept very well. The bathroom was enormous but the shower was quite small.

Figure 2, shows a screen shot from the eye-tracking software with a series of fixations overlaid on the text (the words of the text are shown below the screen shot). It is worth noting that the fixations are not evenly spaced with saccade lengths varying, nor do they fall in an absolutely horizontal line. The fixation dots vary in size which is representative of how long the fixation lasted. Longer fixations are represented by larger dots, shorter fixations by smaller dots. It can also be seen that the fixations do not proceed relentlessly forward through the text. Fixations 1-9 all progress forward through the text but fixation 10 (hidden behind fixation 7) regresses back to an area of the text that has already been fixated. Such movements back to earlier parts of the text are called regressions. Rayner et al. (2012) suggest that short regressions within the current sentence represent word recognition problems, whilst longer regressions back to previous sentences are likely to represent comprehension difficulties. Rayner (2009) reports that such short, sentence level regressions account for the majority of regressions.

Eye-tracking research has revealed that fixation duration, saccade length and rate of regression are a product of a combination of the proficiency of the reader and the difficulty and predictability / ambiguity of the text. This concept is illustrated in Fig. 3.

Figure 3

The Effects of Reader Proficiency, Textual Complexity and Lexical Ambiguity on Eye-Tracking Measures



2.3.1 Reader proficiency

Fixation durations tend to be longer and saccade lengths shorter in less proficient readers than for more proficient readers. Thus, as reading ability improves, shorter fixation and longer saccade lengths emerge as readers process the information from the fixations more quickly and advance further with each with each saccade (Everatt, Bradshaw & Hibbard, 1998). Holmqvist et al. (2011) shows that the number of regressions made also decreased as a function of improved reading skill. However, it should be noted that Hyönä and Nurminem (2006) concluded that long regressions (look-backs) helped readers to develop text-level representations.

2.3.2 Text complexity

As texts become more challenging, fixation length and the number of regressions tend to increase, while saccade length tends to decrease (Blanchard et al., 1989; Jacobson & Dodwell, 1979). A common measure of text difficulty is word frequency and, as might be expected, lower frequency words, that are less familiar, tend to attract longer fixation durations (Williams & Morris, 2004).

2.3.3 Lexical ambiguity / predictability

Semantically ambiguous words, such as bank (associated with both financial contexts and geographic features e.g. river bank) or words with multiple possible pronunciations (e.g. polish, minute or wind) also tend to lead to longer fixation times when the surrounding text fails to fully establish which pronunciation is called for (Sereno, O'Donnell & Rayner, 2006).

The predictability of text also has a role to play. Words which can be readily predicted from the preceding text tend to attracted shorter fixation durations. (Ashby, Clifton & Rayner, 2005; Rayner, Ashby, Pollatsek & Reichle, 2004). Accordingly, longer fixation durations are likely to occur on words which are implausible in light of the preceding text (Rayner, Warren, Juhasz & Liversedge, 2004).

2.3.4 Patterns of fixations

Regardless of textual features, the suggestion that people read at a constant rate, affording consistent attention would be inadequate. Rather, as outlined in Khalifa and Weir's (2009) model, people read for different purposes and this has a significant impact on the way the reader engages with the text. Searching for an advertisement in a local paper, reading a novel and reading an academic article in order to write an essay, are all forms of reading but eye-tracking data from the same reader would show significant differences not just in terms of fixation duration, saccade length and regression rate, but also in terms of the patterns of fixations in relation to the text. Therefore, eye-tracking has an important role to play in helping researchers understand how readers process text according to their different purposes for reading. Central to reading test validity is an awareness of the type of reading elicited by different test items and an understanding of what differentiates the reading processes of successful and unsuccessful test-takers.

2.4 Use of eye-tracking for test validation

Whilst eye-tracking has been used to investigate reading for over 100 years (Rayner, 1998), it is only more recently that it has been deployed for language test validation purposes. Prior to 2010, the majority of eye-tracking research centred upon careful reading "when comprehension is proceeding without difficulty and the eyes are continuing to move forward along a line of text" (Reichle, Warren & McConnell, 2009, p.9). Only more recently has research begun to consider how different types of reading might be deployed, particularly with regard to reading test validation. Some of the earliest eye-tracking studies which used eye-tracking for large-scale reading test validation were carried out by Bax and colleagues. Bax and Weir's (2012) study investigated the reading processes of participants completing items from the computer-based CAE Reading test (Cambridge C1 Advanced) and successfully demonstrated that eye-tracking, using in conjunction with retrospective recall, could be used to infer the different cognitive reading processes, as outlined in Khalifa and Weir's (2009) model of reading, employed as participants answered different test items.

Building on this early use of eye-tracking for test-validation purposes, Bax (2013) investigated the cognitive processes of readers as they completed items from the onscreen version of the IELTS reading test. This study also drew on the work of Weir, Hawkey, Green and Devi (2009), which had used retrospective recall to cast light on the reading processes of test-takers completing the IELTS Reading Test.

The research questions for Bax (2013) centred upon the usefulness of eye-tracking as a methodology for revealing the cognitive processes of test-takers and the differences between successful and unsuccessful test-takers. The findings demonstrated "significant differences between successful and unsuccessful test-takers on a number of dimensions, including their ability to read expeditiously" (Bax, 2013:2). In addition, this study made an important contribution to the field of reading test validation,

demonstrating convincingly that eye-tracking could be used to confirm whether specific cognitive processes were used by test-takers on different test tasks.

However, this research had two particular limitations. The first was that all the participants were Malaysian, and participants from other language backgrounds may have different eye-movement patterns; the second was that the reading processes investigated were confined to forms of local reading i.e. reading a sentence level. Global forms of reading, where connections are made across multiple sentences, were not part of Bax's study.

Bax's (2015) study sought to remedy the limitations of the earlier study by repeating the research using participants with a range of language backgrounds and with the intention of investigating both local and global reading. Once again, the eye-tracking data was triangulated with stimulated recall data. The study also concluded that eye-tracking combined with retrospective recall was a valuable tool for revealing the differences between successful and unsuccessful candidates in terms of some of the reading stages identified in Khalifa and Weir's (2009) model of reading, despite some differences between the Malaysian cohort and the mixed language cohort on some questions. Bax's attempt to investigate global reading (careful or expeditious reading that continues across a number of sentences) was inconclusive, revealing that the reading patterns of participants were incredibly varied and no clear patterns emerged. Nevertheless, once again Bax had successfully applied eye-tracking for reading test validation purposes to great effect, revealing the complexity and variety of cognitive processes engaged in by test-takers.

Brunfaut and McCray (2015) also used eye-tracking and retrospective recall to investigate the cognitive processes of test-takers taking the Aptis reading test. This study used a range of eye-tracking measures to investigate how test-takers reading processes altered as questions became more difficult (and thus whether test-takers engaged in the types of reading intended by the test designer) and how the reading processes of lower scoring test-takers compared to those of higher-scoring. Brunfaut and McCray (ibid) developed a series of hypotheses about how eye-tracking measures would change in response to both task difficulty, text complexity and test-taker proficiency. The data was triangulated with stimulated recall data. Brunfaut and McCray concluded that eye-movement data was particularly useful for giving insight into lower-level reading processes, whilst stimulated recall data was effective in offering insight into higher-level reading processes.

Latimer (2018) used eye-tracking to examine the reading activity of 30 students on an academic reading-into-writing test task. For the first time, the study categorised the eye-tracking data according to the patterns formed by fixations and was thus able to identify the amount of careful reading compared to the amount of selective or expeditious reading. It was concluded that only 30% of the reading undertaken on the task was careful reading (the remaining 70% was accounted for by different forms of expeditious reading) and that higher scoring participants were much more effective at identifying the most relevant parts of the reading texts. This study underscored the importance of expeditious forms of reading for academic study and suggested that identifying the most relevant parts of the text was a key factor in successfully completing academic reading-into-writing tasks.

Bax and Chan (2016) used eye-tracking for another reading test validation study, this time on the General English Proficiency Test (GEPT) developed for Taiwanese learners of English. This research also concluded that eye-tracking data provided a considerable degree of insight into the reading processes of test-takers taking the GEPT and, in line with the literature reported above (Bax, 2013, 2015; Brunfaut & McCray, 2015; Latimer, 2018) concluded that the gaze behaviour of successful and unsuccessful L2 test-takers differed considerably.

The research discussed above outlines how eye-tracking has been adopted by a range of well recognised language tests to map the reading processes of test takers as they complete different test-tasks onto the reading processes outlined in Khalifa and Weir's (2009) model of reading. Eye-tracking, when combined with retrospective recall, can offer nuanced insight into test-takers' reading. Stimulated recall can reveal how test-takers' reading goals are formed and develop. The eye-tracking data can then

reveal how those goals are manifested in different types of reading which facilitate the readers' progress through lower-level reading processes such as parsing onto higher level processes such as building a mental model.

Online reading tests open up opportunities for better learning and assessment experiences when prompt individualised feedback can be provided to direct more effective teaching and learning. However, in order to make feedback as useful as possible, a clear understanding of how participants of different proficiency levels perform the test and how cognitive processes vary from one question to another is necessary. Such information also assists test developers and teachers to raise students' awareness and understanding of the range of cognitive processes and strategies involved in the execution of different linguistic tasks. A deeper understanding of how these skills can be utilised in effective strategies may improve students' meta-cognitive learning and lead to improved comprehension and confidence. Therefore, the aim of this study is to investigate B1, B2 and C1 test takers' cognitive processes as they complete an online reading test.

3 Research Questions

1. Are there differences between eye tracking measures of B1, B2 and C1 test takers on an online reading test?

It is anticipated that the improved reading proficiency of the CEFR (Common European Framework of Reference for Languages) C1 group will be reflected by fewer fixations overall, shorter mean fixation duration, longer saccades, fewer within-word regressions but possibly, a greater number of regressions across multiple words/sentences as more proficient readers build a text level representation.

2. What reading processes and strategies do B1, B2 and C1 test takers report to complete the reading task?

It is anticipated that more proficient readers from the C1 group will make more references to high-level reading processes such as creating a text level understanding, building a mental model (integrating new information) and possibly inferencing. This is because well-developed reading skills should enable more proficient readers, with well-automated low level reading skills, to focus on building a text-level understanding and to consciously apply different types of reading (as set out in Khalifa and Weir's (2009) Goal setter: e.g. careful/expeditious) strategically to find the most relevant areas of the text and focus upon these.

4 Methods

A mixed-methods approach utilising eye tracking and stimulated recall interviews was used to investigate the aforementioned research questions. Eye tracking technology was used to record test takers' eye movements as they completed the reading task. Although eye tracking is a powerful tool to record participants' eye movements on the tasks, eye tracking data are limited in revealing test takers' intention of certain reading behaviours. Individual stimulated recall interviews and a retrospective questionnaire were used to assist participants to report their goals and reading processes they completed the tasks.

4.1 Participants

The participants were 54 English as a Second Language (ESL) learners who were studying at British

universities. Their ages ranged between 18 and 45 years old, see Table 1. Regarding the gender, 44.6% were male and 55.4% female. They were from various L1 backgrounds including Chinese, Vietnamese and Romanian. For data analysis, the participants were divided into three proficiency groups – CEFR B1 (n=18), B2 (n=17) and C1 (n=19), see Table 1. The difference observed was significant, $F(2, 54) = 179.50$, $p<.001$. Post hoc tests showed that differences between all individual pairs were significant at the 0.05 level.

Table 1
Mean IELTS Reading Scores of Test Takers

	B1 (n=18)		B2 (n=17)		C1 (n=19)	
	M	SD	M	SD	M	SD
IELTS reading score	5.21	0.46	6.02	0.65	8.42	0.59

4.2 The task and reading passage

The study used the reading module of the General version of the *Linguaskill* test which is an online, multi-level test designed to assess candidates' English proficiency in daily life. The reading task under investigation requires candidates to read a passage of text of 477 words and answer five multiple choice questions. The reading task is targeted at students at B2 or above.

4.3 Test interface

The test was presented on a computer screen. A webpage was developed using a webpage development tool (Wix.com) to facilitate presenting the relevant sections of the *Linguaskill* test to participants. The webpage closely resembled the *Linguaskill* computerised test but the text did not scroll and instead participants needed to change page by clicking an onscreen button. Participants selected their answers by clicking onscreen buttons.

4.4 Data collection procedures

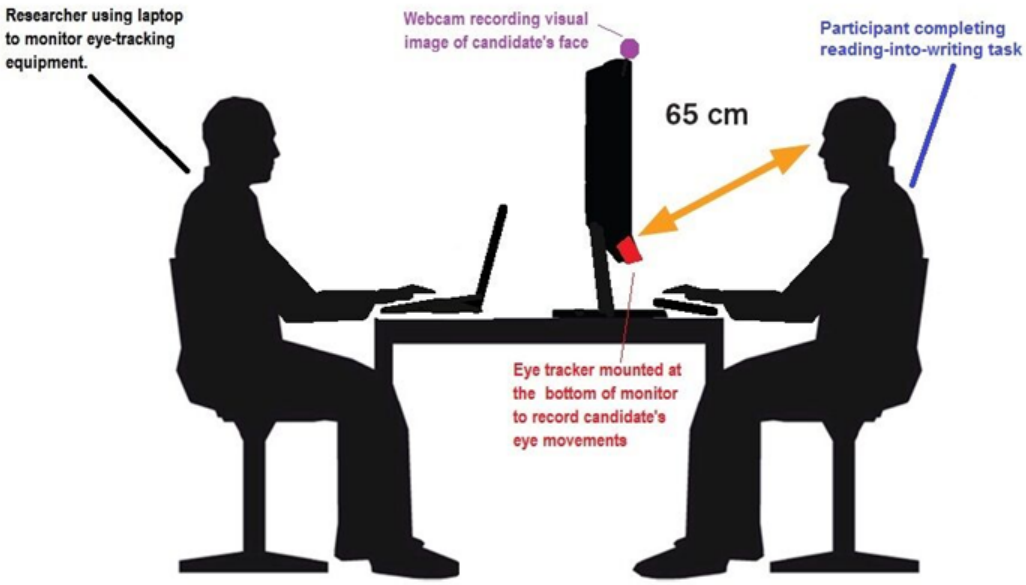
The data collection procedure was conducted with one participant at a time. Participants were shown how to change page and select answers before starting the test. The participants' eye movements on screen were recorded with a Tobii TX60 mobile eye-tracker. The setup of the eye-tracking equipment is illustrated in Fig 4. Immediately after the test event, participants were shown a replay of their test overlaid with their eye-movements (in the form of a moving dot) and asked to recall their processes during the tasks. Each data collection session lasted about 40 minutes, Table 2. The test interface and procedures were first piloted with three participants.

Table 2
Data Collection Procedures for Each Data Collection Session

Procedures	Minutes
Introduction and consent Calibrate eye-tracker and demonstrate functionality of Linguaskill	10
Completion of an extended listening and an extended reading task	15
Retrospective stimulated recall interview – audio recorded	10
Questionnaire	5
Total	40 minutes

The arrangement of the eye-tracking equipment is illustrated in Fig. 4.

Figure 4
Arrangement of Eye-tracking Equipment



5 Data Processing and Analysis

The eye tracking data was collected using a Tobii X2-60 eye tracker which takes 60 reading per second and has an accuracy of 0.4 – 0.6 of a visual degree and precision of 0.34 – 0.74 of a visual degree at a distance of 450mm – 800mm (Tobii, 2016). The range of participants' movements, towards or away from the monitor fell within these distances. Data was collected from 65 participants, but only data from 54 participants (B1: n=18; B2: n=17; C1: n=19) were included in data analysis. This is because the eye tracking data from some participants was not deemed sufficiently accurate (e.g. on occasion reading glasses can distort the eye tracking signal) or there were gaps in the recording where the eye-tracker lost contact with the participants' eyes. The raw data for this research was processed using the Tobii I-VT fixation filter. For full details of the algorithm, please refer to Olsen (2012).

A total of 51 Areas of Interest (AOIs) were identified from the eye-tracking area. Areas of interest are specific areas of the screen where a box, usually a small rectangle, is drawn around a word, phrase or sentence. Every fixation that occurs within the rectangle or shape is tagged with the name of the AOI. An AOI was identified for each reading question and answer option. Each sentence of the reading passage was also given a separate AOI.

As mentioned in the literature review section, eye tracking data is divided into fixations and saccades. For each fixation, the eye tracking software provides the exact location on screen of the fixation (in the form of a coordinate in screen pixels), the time elapsed since the start of the recording (in milliseconds) and the duration of the fixation (in milliseconds). To answer RQ1, eight different eye-tracking measures were used (see RQ1). Potential differences in those eye-movement measures between test takers at different CEFR levels were examined by means of one-way ANOVAs with the different eye-tracking measures as dependent variables and the three levels of test takers as the between subjects' independent variables. Most assumptions were met. However, the data for three of the eight measures were not normally distributed. The results were compared using a non-parametric test but this did not change the results. As One-way ANOVA can be considered robust to non-normality (Maxwell & Delaney, 2004)

when the sample sizes (numbers in each group) are equal, or nearly equal (Lix, Keselman & Keselman, 1996), we decided to report the results. One-way ANOVAs were followed by subsequent pairwise comparisons to compare measures between CEFR-level pairs.

To supplement the eye tracking data, individual retrospective recall interviews were conducted and audio recorded. The recordings were transcribed. The transcripts were reviewed question by question to aid interpretation of the eye-tracking data. The data was analysed through both deductive and inductive approaches (Yin, 2011). From the aforementioned literature on reading processes and an initial read through the transcripts, two themes were identified: levels of processing and test-taking strategies. 274 transcript segments were identified. Specific features were then identified through an inductive approach by reading the transcripts carefully several times, and coding was carried out using NVivo 12. A list of 10 categories emerged from the analysis, see Table 3. This process included several rounds of discussion and sample coding by the two researchers. After the coding scheme was finalised, one researcher coded all segments. 50% of the segments were double coded by the other researcher. The agreement rate was 93.6%, with discrepancies noted, discussed and resolved. The coding scheme, with examples of a coded segment from each task is provided in Table 3.

Table 3

Retrospective Recall Coding Scheme

Category	Examples
Integrating information	… because when the guest arrived…(reading from text) "the car park wasn't very big and I was worried it would be full with other guests' cars"…but, on the other hand he said "the porter came out, said good evening and showed me where to park" (P22)
Evaluating their own understanding or progress	I don't understand "she wasn't shown to the table" (P14)
Making inferences of the content	She was extremely satisfied, she really liked it. (P16)
Establishing representation at levels of clauses and sentences (e.g. by quoting clauses/ sentences from the passage)	She said "the bed looked old but the mattress was not too hard or too soft" (P42).
Analysing the question	The question asked what the receptionist was saying. (P08)
Eliminating distractors	I excluded bathroom, because it wasn't about how clean the bathroom (P12)
Guessing	Maybe. So I guessed, I guessed option B (P31)
Linking content to a question or an answer option	Because in that para, there are here [pointing to the screen]…it shows the bathroom was enormous but the shower was small, and I think it might be this option (P40)
Reviewing the options	So first option is not given, the second one is wrong so it must be the third one (P27)
Predicting the next content or location of answer	Mostly the question comes after one paragraph so there was this one short paragraph and I was expecting there would be a question about this (P47)
General procedures – alluding to the test taking procedure they went through	First of all I read the paragraph and try to understand the words say and I will move past them and try to find the correct one. (P25)

In addition, a short questionnaire was used to assist participants to report their primary reading goals for each question. Khalifa and Weir's (2009) model argues that L2 learners' reading processes are influenced by their goals. The goal setter decides what type of reading to use (See Fig. 1). An important distinction between higher-level and lower-level learners is that the former are skilled in setting appropriate goals

to achieve the reading needs set by a language task. Participants were asked to select ONE reading goal they had on each question. The options include:

1. searching for the relevant parts in the text
2. searching for exact words which matched the question
3. scanning the whole text quickly
4. reading the whole text carefully
5. reading some sentences carefully
6. linking information from different parts of the text
7. using other strategies

6 Results

6.1 Results in relation to RQ1: differences in eye-tracking measures between B1, B2 and C1 groups

Research question 1 relates to whether, as reading proficiency improves, readers would generate fewer fixations overall, shorter mean fixation durations, longer saccades, fewer within-word regressions but possibly, a greater number of regressions across multiple words/sentences (as they build a text level understanding).

First, all participants' performances on the reading task are provided in Table 4. As shown in the mean scores, the participants in the higher proficiency groups tended to perform better than the lower proficiency groups but the difference was small. The difference observed was significant, F (2, 62) = 8.199, p<.001. However, post hoc tests showed that differences between B1 and B2 were non-significant. Most participants on average scored 4 or above out of 5 on the reading task with several B1 and B2 participants reporting "actually I didn't struggle with any of them" (P09, B2). This might indicate that the reading task was not challenging enough. The implications for reading test design will be discussed in the Discussion section.

Table 4

Participants' Performances on the Test Tasks

	B1 (n=18)		B2 (n=17)		C1 (n=19)	
	M	SD	M	SD	M	SD
Reading score	4.00	0.95	4.12	1.05	4.95	0.23

From the eye-tracking measures, shown in Table 5, it can be seen that some of the measures meet the patterns expected, whilst others do not. To remind the reader, it was expected that the number of fixations, overall fixation duration, mean fixation duration and within word regressions would decrease as proficiency increased whilst saccade length and the proportion of regressions across words/sentences would increase. Although the C1 test-takers made the fewest fixations and had the lowest fixation time (both on fixations on the entire task screen and on the reading passage), unexpectedly, the B2 test-takers made more fixations and took more time fixating on the task than the B1 test-takers, see Table 5.

In addition, the mean fixation duration does not follow the expected pattern. B1 and C1 test-takers had the same mean fixation length (B1: M=166.58 milliseconds, SD=103.45; C1: M=166.72, SD=89.34) while B2 test-takers' mean duration per fixation was 181.41 milliseconds.

Table 5

Comparison of Eye-tracking Data between Groups

	B1 (n=18)		B2 (n=17)		C1 (n=19)		One-way ANOVA		Pairwise
	M	SD	M	SD	M	SD	F.	p	
Number of fixations on screen during the reading test	744.33	233.83	819.76	378.80	500.79	198.40	6.72	.003	B1-C1* B2-C1*
Total fixation time (in seconds) on screen during the reading test	124.41	48.85	148.05	83.39	83.52	37.34	5.572	.006	B2-C1*
Duration per fixation (in milliseconds)	166.58	103.45	181.41	105.70	166.72	89.34	98.557	.000	B1-B2* B2-C1*
Number of fixations on reading passage	471.67	167.47	487.18	218.95	322.00	131.97	5.018	.010	B1-C1* B2-C1*
Total fixation time (in seconds) on reading passage	78.88	35.46	86.39	46.63	53.29	24.44	4.200	.020	B1-C1* B2-C1*
Average saccade length on the reading text (in number of character spaces)	7.97	1.40	9.15	2.30	11.48	2.71	12.051	.000	B1-C1* B2-C1*
Proportion of within-word regressions on the reading text	17.55	4.10	16.55	4.46	15.44	5.07	.980	.382	N/A
Proportion of regressions along the same sentence or across sentences	5.79	2.04	5.67	2.04	7.51	2.72	3.677	.032	B2-C1*

Note: *significant difference obtained in the pair-wise comparison

Nevertheless, other measures including *average saccade length* and *proportion of within word regressions and proportion of regressions across words or sentences* suggest that the higher level test-takers might have found the reading passage less cognitively demanding. The post hoc pairwise comparisons show that for saccade length, differences between B1-C1 and B2-C1 were significant, whilst for regressions across words/sentences only the differences between B2 and C1 groups were significant. This interpretation is also supported by the proportion of within-word regressions. The lower-level groups had a higher proportion of within-word regressions which often indicates difficulty at word-level decoding, though the differences observed between proficiency groups were non-significant. The retrospective recall data suggested that decoding may have caused the lower-level test-takers some problems with one B2 test-taker (P07) admitting to finding the language "a little bit confusing" and "trying to match" words from the questions/answer options with words in the passage. Another B2 test-taker (P15) admitted "I'm not really confident about "complex". I really don't know what is that". A B1 test-taker (P04) admitted "I don't know this (word) '*outside*'".

Overall, the data relating to proficiency suggests that the C1 test takers were able to process the reading passage from the test with a much higher processing efficiency than the lower-level test takers. However, to understand what might have led to the unexpected results in relation to the high number of fixations and the fixation time for the B2 test-takers, the retrospective recall data was examined. From the retrospective recall data, it emerged that the B2 participants may have adopted a cautious approach to taking the test, double checking answers in an attempt to ensure they got the answers correct. For example:

Example 1: P07 (B2) I was taught that you read the passage and then go back to read the questions and then the passage again, so that's what I did

Example 2: P18 (B2) This a little bit tricky but you can read again and again

Example 3: P23 (B2) I want to make sure the answer is right so I check the two (wrong answers)

Example 4: P39 (B2) It has questions looking like it might be right but I have to…you know, check

This may partially account for why B2 participants took more fixations (and inevitably time)

to complete the task than the B1 participants and had the longest mean fixation duration. A few C1 participants also reported checking answers unnecessarily but perhaps the increased proficiency of this group enabled them to do this quite efficiently, resulting in a smaller effect in terms of increased number of fixation and fixation time on task. The B2 test-takers mean duration per fixation may also have been artificially increased if B2 test takers deliberately read more slowly and carefully in response to a test situation.

To summarise the results for RQ1, differences did emerge from the eye-tracking data that suggested that the C1 test-takers were able to process the text on the test more efficiently than the lower-level test-takers. However, the eye-tracking results for the B2 test-takers did not generally conform to expectations from the literature, the retrospective recall data seemed to suggest this was because the B2 test-takers adopted a very cautious approach to taking the test.

6.2 Results for RQ2: what levels of processing and strategies do B1, B2 and C1 test takers report to complete the reading task?

Research question 2 relates to whether more proficient readers from the C1 group would make more references to high-level reading processes such as creating a text level understanding, building a mental model (integrating new information) and possibly inferencing. The rationale behind this was that more proficient readers, with well-automated low level reading skills, would be able to focus more on building a text-level understanding and to consciously apply different types of reading strategically to find the most relevant areas of the text and focus upon these.

The retrospective recall data was coded. Inevitably, there were differences between the number of coded comments for each group, therefore in order to compare groups, the number of comments were standardised to percentages. The data is presented in Table 6.

Table 6

Results of Retrospective Recall Data Analysis

Retrospective recall data	B1 group	B2 group	C1 group
Levels of processing	%	%	%
Creating text level representation	0	0	0
Integrating information	1.25	0	0
Making inferences of the content	17.5	18.37	15.63
Evaluating own understanding or progress	10.64	12.24	10.94
Establishing representation at levels of clauses and sentences (e.g. by quoting clauses/sentences from the passage)	6.38	12.24	1.56
Test taking strategies			
Analysing the question	0.00	4.08	0.00
Eliminating distractors	6.78	8.16	14.06
Guessing	2.13	2.04	0.00
Linking content to the question options	21.28	18.37	17.19
Reviewing answer options	17.02	10.40	12.50
Predicting the next content or location of answer	2.13	2.04	7.81
General procedures – alluding to the test taking procedure they went through	14.89	12.24	20.31

Regarding level of processing (see Fig 1), references to inferencing were most common across levels, with the B2 test-takers making more references to inferencing (18.37%) than the B1 (17.50%) or the

C1 participants (15.63%). The other common comments across levels related to evaluating one's own understanding. The B2 participants made the highest percentage of comments regarding establishing representation at clause and sentence level.

There was only one mention of integrating new information (by a B1 test taker), which relates to building a mental model of the text in the Khalifa and Weir (2009)'s model of reading. There were no references made to creating a text level representation (which is one of the highest levels of processing) from any of the test taker groups. However, this may be because all reading questions specifically targeted one paragraph of the text at a time. In other words, for the reading task that was investigated, there were no questions which required test-takers to make a judgement about the whole text (for example, asking what the purpose or overall theme of the text was).

It was expected that higher-level test takers would make more references to higher level reading processes such as inferencing. However, when the questions were examined, many questions appeared to rely on inferencing (e.g. linking background knowledge of the world with implicit propositions) and therefore all the participants mentioned inferencing. If the test had contained a longer passage and had included questions which required test-takers to make more subtle inferences across bigger sections of text perhaps the C1's would have been more successful at this.

Even though all the test-takers scored quite highly, inevitably, some questions proved relatively more difficult. Table 7 below shows the scores achieved on individual questions.

Table 7

Reading Scores by Question and CEFR Level

	Mean	SD	Mean	SD	Mean	SD	Mean	SD
CEFR level	B1 (n=18)		B2 (n=17)		C1 (n=19)		All test-takers (n=54)	
Reading Q1	0.78	0.422	0.91	0.294	1.00	0.000	0.89	0.312
Reading Q2	0.91	0.288	0.95	0.213	1.00	0.000	0.95	0.211
Reading Q3	0.74	0.449	0.73	0.456	0.95	0.224	0.80	0.403
Reading Q4	0.83	0.388	0.82	0.395	1.00	0.000	0.88	0.331
Reading Q5	0.74	0.449	0.77	0.429	1.00	0.000	0.83	0.378

To try and understand why some of the questions were more difficult than others, the easiest and most difficult questions were compared. Both questions required test-takers to link background knowledge of the world with one or two propositions from within a single sentence. Question 3 was the most difficult question for the test takers (see Table 7). The text relating to question 3 was:

> *I was very satisfied when I saw my room. It was nicely decorated with antique furniture and traditional wallpaper. Although the bed looked old, the mattress was not too soft or too hard and I slept very well. The bathroom was enormous but the shower was quite small.*

Question 3 asked "What does the guest say she liked about the bedroom?" with answer options of "The bathroom was clean", "The bed was comfortable" and "The decoration was modern".

Question 2, which was the easiest question, asked the test takers to complete the sentence "*When the guest arrived…*" and the answer options were "*The car park was full*", "*Nobody offered to carry her bags*" or "*There were not enough outside lights*". The text read:

> *I arrived quite late at night. The car park wasn't very big and I was worried it would be full with other guests' cars. But the porter came out, said 'Good evening, madam' and showed me where to park, before helping me with my suitcase. It was quite dark in the car park, with only two lights by the entrance, so it was hard to see where I was going.*

For both questions some test takers appeared to be able to answer the question by linking background knowledge of the world with just one proposition.

Question 3

Example 6 "I slept very well" that means the bed is comfortable I think (B2)

Example 7 "the mattress was not too soft or too hard" so this really help me to know. (B2)

Question 2

Example 8 I see "It was quite dark in the car park", so I know. (B1)

Example 9 There were not enough outside lights, it says here… "only two lights" (B2)

Therefore, it is difficult to speculate why question 2 was easier than question 3, perhaps the ease with which discriminators could be eliminated played a part. The researchers examined the eye-tracking data to see if there was any correspondence between the number of fixations (forward or regressions) that occurred on each question and the question's difficulty but no clear correlations emerged. It is likely that the "over-processing" (double checking of correct answers by eliminating wrong answers even after the answer had been selected) mentioned earlier would prevent any such correspondences from emerging from the eye-tracking data.

After examining the data about levels of processing, the retrospective recall data regarding test taking strategies is now discussed. In terms of test-taking strategies (Table 6), C1 test-takers made the most references to test-taking procedures (20.31% of comments), linking the content of the reading passage to the answer options (17.19% of comments) and eliminating distractors (14.06% of comments). Together these results suggest that the C1 test-takers adopted a strategic approach to taking the test, making connections between the individual questions and the related parts of the passage and double checking their selection of the correct answer by also eliminating the wrong answers as outlined by P56 (C1).

Example 10 *So it has to be comfortable, that (option) B. But still, since it's a test I will read all the options. SO I go to the third option…*

The B2 test-takers made most references to linking the content of the reading passage to the answer options (18.37% of comments), test taking procedures (12.24% of comments) and reviewing the answer options (10.20% of comments). These comments suggest that the B2 test-takers also adopted different strategies, as illustrated in Examples 1-4 above. However, they made less mention of eliminating distractors as a strategy. It seems that the B2 test-takers, whilst adopting a cautious approach, focused on rereading the questions and answer options rather than returning to the reading passage.

The B1 test-takers also made most references to linking the content of the reading passage to the answer options (21.28% of comments), followed by reviewing the answer options (17.02% of comments) and test-taking procedures (14.89% of comments). This group made few references to eliminating distractors. This fact, in conjunction with making a high number of comments about linking the content of the reading passage to the answer options, suggests that most of their attention was absorbed by trying to discern the correct answer and they did not report double checking their answers by eliminating distractors as the two higher-proficiency groups did.

Moving onto the questionnaire data (Table 8), the results suggest that the C1 test-takers engaged in more selective reading. This is suggested by 35.96% of C1 responses reporting reading some sentences carefully and 30.34% of responses reporting searching for the relevant parts of the text. In contrast just 2.25% of C1 responses reported reading the whole text carefully.

Although the lower level test-takers also reported reading some sentences carefully (B1:29.47%, B2:22.73%) and searching for the relevant parts of the text (B1:17.89%, B2:22.27%), they reported searching for exact words which matched the question (B1:21.05%, B2:19.09%) and reading the whole text carefully (B1:7.37%, B2:12.73%) more than the C1 participants.

In combination, these figures suggest that whilst all the test-takers engaged in goal-setting (as per Khalifa and Weir's 2009 model of reading) to some extent (for example: singling out parts of the reading passage most relevant to the current question), the C1 test-takers did this more often by applying careful reading more selectively and using expeditious reading to search for the most relevant sections more frequently.

Table 8

Results of the Questionnaire Data

Goal setting for all reading questions	B1 (n=18 on 5 Qs)	B2 (n=17 on 5 Qs)	C1(n=19 on 5 Qs)
	%	%	%
Searching for the relevant parts in the text	17.89	27.27	30.34
Searching for exact words which matched the question	21.05	19.09	15.73
Scanning the whole text quickly	9.47	7.27	7.87
Reading the whole text carefully	7.37	12.73	2.25
Reading some sentences carefully	29.47	22.73	35.96
Linking information from different parts of the text	12.63	10.00	4.49
Using other strategies	2.11	0.91	3.37

To remind the reader, using retrospective recall and questionnaire data, RQ2 investigated which reading processes and strategies B1, B2 and C1 test takers reported. The results suggested that the task elicited inferencing and evaluating own understanding from all three groups of test-takers. B2 test takers made more reference to establishing representation at the level of clauses and sentences than the other groups. Whilst all the questions required test-takers to use inferencing, none of the questions seemed to elicit higher level processes targeting at the whole-text level understanding.

In terms of reading strategies, the findings showed that both the C1 and B2 test-takers adopted a strategic approach by, for example, spending time reading specific parts of the text and eliminating distractors. The C1 test-takers appeared to deploy different types of reading more deliberately and more effectively than the lower level groups. The B2 test-takers seemed to have adopted a cautious approach to the test, often rereading text carefully. The B1 test-takers, on the other hand, focused more on linking the content of the reading passage to the answer options, suggesting that most of their attention was absorbed by trying to discern the correct answer.

7 Discussion and Conclusion

This study has several limitations. The first of which is the small sample size of 54 participants completing an online reading test of five questions. Gathering eye-tracking data is time consuming as it can only be gathered from one participant at a time. The small sample size may mean that results are not generalisable. The type of reading questions (i.e. multiple-choice questions which focused mostly at the levels of clauses, sentences and single paragraph) examined was limited. This might have impacted the range of reading processes employed by the test takers. In addition, eye-tracking data can only be considered indicative of reading; it cannot record the cognitive or reading processes that are taking place. However, with the help of retrospective recall data and a retrospective questionnaire, such processes may be inferred.

This study hypothesised that the higher reading proficiency of the C1 group would be reflected by fewer fixations overall, shorter mean fixation duration, longer saccades, fewer within-word regressions but possibly, a greater number of regressions across multiple words/sentences. These expectations were partially met. The C1 group did report fewer fixations and less fixation time on task and on the reading

passage. They also had longer saccades, fewer within word regressions and more regressions across words/sentences on the reading passage. However, the B2 group confounded several of these expectations when compared to the B1 group. The retrospective recall and questionnaire data suggested that both the B2 and C1 groups adopted a strategic approach, spending time reading and eliminating distractors rather than relying solely on identifying the answer by understanding the text. In addition, the B2 test-takers seemed to have adopted a cautious approach to the test, often rereading text carefully, which perhaps accounts for the unexpected profile of the B2 eye-tracking data. The B1 test-takers seem to have mainly focused in discerning the correct answer, rather than trying to eliminate distractors. In terms of reading processes, the retrospective recall and questionnaire data suggested that the C1 test-takers deployed different types of reading more deliberately and more effectively than the lower level groups.

In relation to these results this study draws conclusions and makes recommendations in relation to (1) reading items investigated, (2) the multiple-choice format more generally and (3) eye-tracking as a methodology.

7.1 Reading items investigated

The online reading test functioned reasonably well in terms of eliciting a range of inferencing skills from participants and the questions distinguished between all three levels, although the difference in test scores between B1 and B2 test-takers was quite small. However, considering that the test was targeted at students at B2, the test did not appear to elicit some of the higher-level reading processes including building a mental model or building a text level representation. Adding questions targeted at these higher level processes may improve differentiation between B2 and C1 levels. It should also be noted that both the B2 and C1 groups seemed to engage in over-processing of the text, taking time to reread the text and, in some cases, to spend time eliminating distractors rather than relying on an understanding of the text to identify the correct answer. Whilst this test was aimed at B2 and therefore over processing by the C1 test-takers might be expected, this once again suggests that the addition of some more cognitively demanding, text-level questions might improve the cognitive representation of the test.

7.2 Multiple-choice format reading tests

In relation to the multiple-choice format more generally, the results suggest that, in some cases, multiple-choice questions can encourage test-taking strategies such as discounting the distractors rather than building a broader understanding of the text. The results also suggest that if individual multiple choice questions are linked to small sections of text, there is limited opportunity for test-takers to use the types of reading often demanded in life beyond the test such as expeditious reading skills (skimming/scanning/searching). Several suggestions are made as potential measures to combat this:

1. Making reading passages longer to discourage test takers from reading all sections carefully. This would potentially encourage test-takers to use expeditious reading such as searching / scanning in order to locate the sections which then need to be read carefully in order to answer questions. This would require the inclusion of some surplus information which is not the focus of any questions.
2. Manipulating the time allowed to complete tests could have an influence on reading behaviour. For multiple choice questions, even competent readers, as shown by the C1 participants in this study, would be tempted to spend time eliminating distractors instead of relying on their understanding of text to select the correct answer. Therefore, ensuring test-takers have only just enough time to complete some sections of the test (or perhaps even insufficient time to complete all the questions in that section) may force test takers to deduce the answer rather than using ruling out distractors. Alternatively, online reading tests could require higher-level test takers to answer more items in the same time limit.

3. In order to ensure that high level reading skills, such as creating a text-level representation and intertextual representations, are sampled, some questions need to target ideas/themes conveyed across an entire text or multiple texts in the case of intertextual representation.

In conclusion, in order to avoid underrepresentation of the target reading processes, test designers should clearly identify which reading processes they expect test takers to use when answering questions and should try to ensure that different questions target different reading processes, especially those requiring expeditious reading and those requiring understanding at the whole text level.

7.3 Eye-tracking as a research methodology

This study has shown that eye-tracking can provide a useful record of reading activity; however, it has also shown that without retrospective recall and retrospective questionnaire data to aid interpretation, very few conclusions can be drawn if the data does not "fit" typical expectations and important insights may be overlooked.

The researchers would also caution that eye-tracking data is time consuming to collect as it can usually only be collected one participant at a time. This makes it difficult to collect enough data to overcome individual test taker differences that frequently emerge from eye-tracking data.

It should also be noted that often, when capturing eye-tracking data, the text on screen cannot be scrolled because analysis of some eye-tracking metrics such as saccade length, usually rely on screen coordinates. Text also needs to be suitably spaced to allow for clear discrimination between lines of text. This means that screen shots or test functionality may need to be adapted.

A modest amount of eye-tracking generates a huge amount of data which can be time consuming and complex to analyse. However, as demonstrated in this study, with a careful research design integrating retrospective recall and questionnaire data, it is possible to investigate readers' goal setting and which types of reading (as per Khalifa and Weir's model) have been used on a reading test. It is hoped that the research methods demonstrated would be useful for researchers who share similar interests in examining test takers' reading processes. Future studies should further explore test takers' reading processes on different item formats.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732935

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurements in Education (NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Ashby, J., Rayner, K., & Clifton Jr, C. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *The Quarterly Journal of Experimental Psychology Section A*, *58*(6), 1065-1086. https://doi.org/10.1080/02724980443000476

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

Bax, S. (2013). *Readers' cognitive processes during IELTS reading tests: Evidence from eye-tracking*. British Council, ELT Research Papers, 13–06.

Bax, S. (2015). *Using eye-tracking to research the cognitive processes of multinational readers during an IELTS reading test*. IELTS Research Reports Online.

Bax, S., & Chan, S. H. C. (2016). *Researching the cognitive validity of GEPT high-intermediate and*

advanced reading: An eye tracking and stimulated recall study. Language Training and Testing Center (LTTC).

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

Beard, R. (1972). *Teaching and learning in higher education*. Penguin.

Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics*, *25*, 133–150. http://dx.doi.org/10.1017/s0267190505000073

Blanchard, H. E., Pollatsek, A., & Rayner, K. (1989). The acquisition of parafoveal word information in reading. *Perception & Psychophysics*, *46*(1), 85-94. https://doi.org/10.3758/BF03208078

Brysbaert, M., & Vitu, F. (1998). Word skipping: Implications for theories of eye movement control in reading. In Underwood, G. (Ed.) *Eye guidance in reading and scene perception* (pp. 125-147). Elsevier Science Ltd. https://doi.org/10.1016/B978-008043361-5/50007-9

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

Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (2018). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition*, *40*(2), 341–366. https://doi.org/10.1017/S0272263117000109

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179-197. https://doi.org/10.1037/0033-2909.93.1.179

Everatt, J., Bradshaw, M. F., & Hibbard, P. B. (1998). Individual differences in reading and eye movement control. In G. Underwood (Ed.) *Eye guidance in reading and scene perception* (pp. 223-242). Elsevier Science Ltd. https://doi.org/10.1016/B978-008043361-5/50011-0

Glaser, R. (1991). Expertise and assessment. In M. C. Wittrick & E. L. Baker (Eds.), *Testing and cognition* (pp. 17–30). Prentice Hall.

Grabe, W., and Jiang, X. (2014). Assessing reading. In Kunnan, A. J. (Ed.), *The companion to language assessment volume 1: Abilities, contexts, and learners* (pp. 185- 200). Wiley Blackwell.

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

Hyönä, J., & Nurminen, A. M. (2006). Do adult readers know how they read? Evidence from eye movement patterns and verbal reports. *British Journal of Psychology*, *97*, 31–50. https://doi.org/10.1348/000712605X53678

Jacobson, J. Z., & Dodwell, P. C. (1979). Saccadic eye movements during reading. *Brain and Language*, *8*(3), 303-314. https://doi.org/10.1016/0093-934X(79)90058-0

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

Latimer, N. (2018). *Reading during an academic reading-into-writing task: An eye-tracking study* [Unpublished PhD dissertation]. University of Bedfordshire.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research*, *66*, 579-619. https://doi.org/10.2307/1170654

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective (2nd ed.)*. Psychology Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (3rd edition)* (pp. 13–103). McMillan.

Olsen, A. and Matos, R. (2012). Identifying parameter values for an I-VT fixation filter suitable for

handling data sampled with various sampling frequencies. In *Proceedings of the symposium on eye tracking research and applications* (pp. 317-320). Association for Computing Machinery.

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

Rayner, K. (2009). The 35th Sir Frederick Bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*(8), 1457-1506. https://doi.org/10.1080/17470210902816461

Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the EZ Reader model. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(4), 720-732. https://doi.org/10.1037/0096-1523.30.4.720

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

Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(6), 1290-1301. https://doi.org/10.1037/0278-7393.30.6.1290

Reichle, E., Warren, T., & McConnell, K. (2009). Using E-Z Reader to model the effects of higher-level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, *16*(1). 1–20.

Sereno, S. C., O'Donnell, P. J., & Rayner, K. (2006). Eye movements and lexical ambiguity resolution: Investigating the subordinate-bias effect. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(2), 335. https://doi.org/10.1037/0096-1523.32.2.335

Tobii, A.B. (2016). *Tobii Studio Users' Manual*. https://www.tobiipro.com/siteassets/tobii-pro/user-manuals/tobii-pro-studio-user-manual.pdf.

Walczyk, J. J. (2000). The interplay between automatic and control processes in reading. *Reading Research Quarterly*, *35*(4), 554–566. https://doi.org/10.1598/RRQ.35.4.7

Weir, C. J. (1983). *Identifying the language problems of the overseas students in tertiary education in the United Kingdom* [Unpublished PhD dissertation]. University of London.

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

Weir, C. J., Vidakovic, I., & Galaczi, E. (2013). *Measured constructs: A history of the constructs underlying Cambridge English language (ESOL) examinations 1913-2012*. Cambridge University Press.

Weir, C. J., Yang, H., & Jin, Y. (2000). An empirical investigation of the componentiality of L2 reading in English for Academic Purposes. *Studies in Language Testing 12*. Cambridge University Press.

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

Williams, R., & Morris, R. (2004). Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology*, *16*(1-2), 312-339.

Wolverton, G. S., & Zola, D. (1983). The temporal characteristics of visual information extraction during reading. In Rayner, K. *Eye movements in reading* (pp. 41-51). Academic Press.

Yin, R.K. (2011). *Qualitative research from start to finish*. Guildford Press.

Nicola Latimer is a Visiting Researcher at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire and also lectures on the MA Applied Linguistics at Queen Mary University London. Her research interests include integrated assessment, assessing L2 reading and using eye-tracking to investigate reading. ORCID number: 0000-0003-1412-9326

Sathena Chan is a Senior Lecturer in Language Assessment at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire. Her research interests include integrated assessment, L2 language processing (reading, writing and summarising), and the role of technology in language learning and assessment. ORCID number:0000-0002-7852-6737