Using eye-tracking research to inform language test validity and design

1. Introduction

It has been argued that cognitive interpretative claims of any language tests should "not be foregone conclusions, [but] need to be warranted conceptually and empirically" (Ruiz-Primo et al., 2001, p.100). Since Bax's (2013) innovative eye-tracking study to examine test takers' reading patterns on IELTS reading items, there has been a marked increase in language testing research involving eye-tracking. Eye-tracking appears to offer a useful methodology for exploring cognitive validity in language tests (Glaser, 1991), 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 paper reports on a recent study which examined the cognitive validity of two level-specific English Proficiency Reading Tests (corresponding to CEFR[1] B2 and C1) through a mixed-method approach utilising eye-tracking technology, a self-report checklist and retrospective simulated recall interviews. In particular, the study aimed to examine relationship between item types and students' reading processes so as to inform test validity and design.

2.1 Empirical bases to support a componential model of reading for testing

An important question for validating the construct of a reading test is whether reading 'can be broken down into underlying skill components for the purposes of teaching and testing' (Weir, Yang and Jin, 2000, p. 14). Grabe (1991) offers a useful list of reading skill components including automatic recognition skills, identifying main ideas of a text, and inferencing. Grabe and Stoller (2002) classify these reading

[1] Common European Framework of Reference for Languages

processes into higher and lower-level processes, which are discussed in more detail below.

Based upon the view of the componentiality of reading skills in Weir's (2005) socio-cognitive framework for language test development, Khalifa and Weir (2009) specify the construct of reading by closely relating the putative reading components to Kellogg's (1996) model of reading, which was empirically grounded in cognitive psychology.

Table 1. Reading processes (Khalifa and Weir, 2009)

Lower level processes	Word recognition
	Lexical access
	Syntactic parsing
	Establishing propositional meaning at clause and sentence level
Higher level processes	Inferencing
	Integrating information across sentences
	Creating a text level structure
	Integrating information across texts

Khalifa and Weir (2009) argued that it is vital for reading tests offered at different proficiency levels to determine and control the level of these reading processes which students need to use during the tasks.

2.2 Cognitive validation of reading tests: under-representation of higher-level processes

Khalifa and Weir (2009), for the first time in the language testing literature, examined the extent to which different levels of reading processes (see Table 1) were operationalised in the suite of Cambridge English reading examinations. Since then,

a number of studies used a similar approach to examine the cognitive validity of reading tests (e.g., Brunfaut & McCray 2015, Owen 2016, Wu, 2014).

By means of expert judgment and retrospective test taker checklist, Wu (2014) investigated the cognitive processes targeted in the GEPT Reading Tests. Owen (2016) examined the cognitive processes activated by IELTS and TOEFL iBT reading tests. He used video recordings of the student reading to facilitate the verbal recall protocols he employed. The findings provided evidence of the frequent application of (local) expeditious strategies in completing the IELTS reading module (2016, p.144). However, he argued that expeditious reading was not tested separately from careful reading in IELTS, but rather the two types of reading tended to co-occur as students searched quickly for information, which then helped them to answer a question by word matching with the question to locate sentences in the text in which the answer might be found. Owen (2016) concluded that 'some of the higher-level processes in Khalifa and Weir's model were under-represented in both IELTS and TOEFL, including inferential reasoning and forming a text-level representation (p. 366-367). Weir et al's (2000) research for the Advanced English Reading Test (AERT) in China found that although many students in China were performing adequately in careful reading, this did not necessarily translate into proficiency in expeditious reading. They argued that testing careful reading ability alone was insufficient as a relevant and comprehensive measure of adequate English reading skills for academic study.

The results of these studies consistently show that most reading tests focus on testing students' careful reading skills and some expeditious reading at local level. Enright, Grabe, Koda, Mosenthal, Mulcany-Ernt and Schedl (2000), who carried out the TOEFL 2000 revision, argued for including tasks that require processing beyond the level of searching for information and basic comprehension of main ideas in a text. They proposed that some tasks should necessitate an understanding of how information in a text as a whole is connected, and how to integrate information from a variety of texts for use in written assignments. Similarly, AUTHORS (DATE), in their review of research on academic reading, warned that there have been few attempts to develop models for expeditious reading or to include such reading types in research, tests or teaching as compared to the focus on careful reading (p.55).

2.3 Eye-tracking studies to reveal different types of reading processes

Although Khalifa and Weir's (2009) study was influential, their investigation relied heavily on the subjective judgments of experts rather than empirical evidence of how students actually completed the reading tests. Other previous process studies have used think-aloud protocols and interview (e.g. Cohen & Upton, 2007). Although think-aloud protocols provide rich data to reveal students' processes on a language task, the think-aloud procedures inevitably disrupt the processes under investigation (Cooper & Holzman, 1983; Russo, Johnson & Stephens, 1989). Recently, an innovative method to supplement expert judgment and think-aloud protocols has become available through developments in non-intrusive eye tracking technology (Eger, Ball, Stevens & Dodd, 2007).

In his seminal study on eye-tracking and test taking, Bax (2013) examined 38 students' reading processes during an IELTS reading test and reported significant differences between successful students (i.e. who answered the items correctly) and unsuccessful students (i.e. who answered the items incorrectly) in terms of their expeditious reading skills. He found that unsuccessful students were not able to locate the site of a correct answer in the text expeditiously whereas successful students tended to employ conscious strategies to locate relevant information in the text to answer the questions (i.e. *sentence completion* and *matching*). As a result, Bax (2013) concluded that 'if the ability to read expeditiously is an important marker of successful as opposed to unsuccessful readers, then future reading test developers might well choose to give expeditious reading an even more central place in their specifications than they do currently (ibid, p.18)'.

The differences in students' reading processes between successful and unsuccessful task completion were also examined in another eye-tracing and test taking studies. Brunfaut and McCray (2015) examined 25 students' reading processes while completing Aptis reading tasks, and found that successful item completion was most often associated with a careful global and/or local reading approach. Expeditious reading was conducted by only some test-takers on some tasks. Although the students engaged in a wide range of cognitive processes on the Aptis reading tasks, including the lower- and higher-level processes, Brunfaut and McCray argued cautioned that the test did not tap into reading processes at the intertextual level. In addition to differences between students, notable differences in students' reading

processes on different item types were reported. For example, *gap-fill* items elicited relatively more careful local reading, while *sentence-ordering* and *matching headings* items involved relatively more careful global reading and some expeditious reading. This shows the importance of investigating the relationship between reading item types and students' processes to evaluate the effectiveness of particular reading item types and to better inform test design in future.

In the light of these studies, using a mixed-method approach, the current study investigated the impact of item type on students' reading processes by examining students' reading processes while completing different reading item types.

3. Methodology

3.1 Participants

As GEPT is developed for Taiwanese learners (see section 3.2 for more information about the test), we sought collaboration with University of Southampton where there was a large cohort of Taiwanese students at the time of the study. The chairman of the Taiwanese Students Association was contacted to help with recruiting participants. Through the Association, information flyers were delivered to students on campus as well as posted on social media. According to the GEPT-CEFR alignment research (AUTHORS, DATEa; Wu and Wu, 2010; Wu, 2014) the proficiency level students who passed the GEPT Advanced and High-Intermediate reading tests are estimated to be equivalent to IELTS 6.0 and 7.0 respectively. We, therefore, aimed to recruited students with an IELTS reading score of 6.0 or above.

As a result, twenty-four students who met the criteria were recruited. All of them were doing Masters-level programmes from disciplines including Business, Education, Translation and Law. Their mean IELTS reading score was 6.58 (SD=0.54), therefore, the population was similar to students expected to take the GEPT High-Intermediate and Advanced test. Their age ranged from 18 to 21. 65% of the participants were female. Eight students were selected randomly to participate in the stimulated recall interview.

3.2 Reading items

Six reading item types were carefully sampled from the High-Intermediate and Advanced General English Proficiency Test (GEPT) Reading paper (for details see Table 2). The GEPT, developed and administered by the LTTC, targets English learners in Taiwan. This test corresponds to Taiwan's English education framework and aligned to the Common European Framework of Reference for Languages (CEFR), provides institutions or schools with a reference for evaluating the English proficiency levels of their job applicants, employees, or students (LTTC, 2016). The GEPT was deemed most suitable for the current investigation as it has a separate section dedicated to expeditious reading skills, which is a rare feature among high-stakes reading examinations.

Table 2. Reading item types

Levels	Item types	Time	Item selected for detailed eye tracking analysis
Higher-intermediate (B2)	**Cloze** – MCQ (n=7)	14 mins	Q3 (sentence level)
	Reading comprehension MCQ (n=7)		Q13 (paragraph level) Q14 (text level)
Advanced (C1)	**Careful reading** Summary Cloze (n=6)	25 mins	Q19 (text level)
	Expeditious reading -		Q21 (paragraph/text level)

	Heading matching (n=5) Which texts? (n=5)		Q30 (multiple texts)
	Total: 30 items		

A testlet with all the selected items was created using Adobe Acrobat DC to facilitate the eye tracking study, with careful attention paid to making the test experience as close as possible to normal GEPT test contexts. The time allocation for each item was calculated based as a proportion of the time allocation in the original test. Participants' performance was marked by the researchers using the marking scheme provided by the LTTC. For the second part of the analysis, of the eye tracking gaze data, 6 representative items were selected for detailed investigation, as listed in Table 2, in ways explained in Section 3.7.

3.3 Eye tracker

The Tobii X2 Eye Tracker was used to track the test takers' eye movements during the reading test. This device is deemed appropriate to the current research purposes, with a 60 Hz eye tracking rate, free head movement for participants, and high-quality tracking of large gaze angles (up to 36°) (Tobii, 2013).

3.4 Reading processing checklist

Based on Khalifa and Weir's (2009) and Wu's (2014) work, a reading processing checklist was developed to assist participants to report the reading processes they employed immediately after they have completed each reading item type. The checklist covers students' reading goals, types of reading employed and structural level.

3.5 Data collection procedures

Data was collected on a one-to-one basis. After completing ethics procedures and information forms, the participants completed the 30 items on a computer, with the Tobii X2 Eye Tracker recording their eye movements on screen. Immediately after they had completed each item type, they reported their reading processes by using a Reading Process Checklist. Eight students further participated in a stimulated recall interview while viewing video footage of their gaze patterns on the test. The key stages of the data collection included:

1. Participants completed personal information and consent forms;
2. Researchers explained the procedures;
3. Researchers calibrated participant's eye fixations;
4. Participants completed the Test (see Table 2) on a computer;
5. Participants filled in the Reading Process Checklist;
6. 30% of participants (n=8) completed a stimulated recall interview; and
7. Participants viewed the video footage of their test event, with their eye-movements measures. They described their reading goals, processes and/or strategies while watching the video.

3.7 Data analysis

The eye tracking data was analysed both qualitatively and quantitatively. Qualitatively, the segments, i.e. Areas of Interest (AoI) of the question and reading text(s) which test takers read to complete each individual test item were analysed. The cognitive demands (e.g. at the levels of lexical, single sentence, multiple

sentences, single text or multiple texts) of each selected reading item were then coded. Findings are discussed in terms of differences between successful and unsuccessful students (i.e. those who answered the item correctly and those who answered the item incorrectly).

Quantitatively, numerical data on each test taker's eye movements, in terms of fixation duration (the length of time a reader fixated on a section of the text), fixation count (the number of times a reader fixated on a section of the text), visit duration (the length of time a reader remained on a section of the text), and visit count (the number of visits a reader made to a section of the text), were generated by the eye-tracking software. The sections of text selected for analysis varied from item to item, and included text as small as a phrase, a correct or incorrect answer in an MCQ item, or text as large as a whole passage.

The eye-movement data was compared descriptively between the successful and unsuccessful students. The non-parametric Mann-Whitney U tests were also conducted to examine the statistical significance of any differences of the eye movements between the successful and unsuccessful students on an item-by-item basis, when the sample size of both groups exceeded n=5. It is recognised that the Mann-Whitney U test can be used with such small samples (see e.g. Sheskin 2003, Hinton 1995, and in particular the example in Wood, Fletcher & Hughes, 1986, page 188). However, the sample size of one of the two groups was often very small, and therefore the results of the inferential statistics should be interpreted with caution.

In addition, close analysis was applied in more qualitative mode to the detailed gaze patterns of each participant, via the heat map and other tools available in the software, in order to tease out other important patterns which might be missed by purely quantitative approaches.

In order to obtain a more nuanced and fine-grained understanding of how students' reading processes might have impacted their test scores, participants were sorted into four groups according to levels of performance (i.e. low, low-medium, medium-high, high scoring groups - see Table 4 in the Results section). Frequencies of participants' responses in the reading process checklist in relation to reading goals, types of reading and levels of reading were then calculated to provide an overall analysis of the processes elicited by different item types.

Verbal protocols elicited in the stimulated recall interview were transcribed to supplement the self-report processing data. The data on each item types were then compared by different scoring groups (see Table 4 in the Results section).

4. Results

4.1 Students' test performance

The students performed relatively well on the High- Intermediate section of the test overall, and relatively poorly on the Advanced section, the Summary task in particular.

Table 3. Students' test performance

		Mean	SD	Max	Min
Levels	Item types	%	%	%	%
Higher-intermediate (B2)	Cloze – MCQ (n=7)	74.40	13.03	100.00	57.14
	Reading comprehension– MCQ (n=7)	71.43	23.33	100.00	28.57
Advanced (C1)	Summary Fill in the blanks (n=6)	28.13	20.81	75.00	0
	Expeditious reading - Heading matching (n=5)	40.83	24.82	80.00	0
	Expeditious reading - Which texts? (n=5)	40.00	25.82	80.00	0
	Total	53.13	14.82	78.33	30.00

As mentioned in Section 3.1, the participants' reading proficiency level, as measured by IELTS, was homogenous. However, their performance on the different item types in this study indicated that they might have an uneven profile of reading skills, as reported in Weir et al's (2000) study with L2 students in China.

For purposes of the eye tracking analysis, students were divided according to their correct or incorrect answers on a question-by-question basis. To interpret the self-report and interview data, a finer distinction was possible, with participants categorised into either the low-scoring (achieving 0-25% of the total score), low-medium (26-50%), medium-high (51-75%) or high-scoring (76-100%) group, depending on their performance on each task (see Table 4).

Table 4. Grouping based on participants' test performance

	Cloze	Reading comprehension	Summary	Expeditious reading
Low (0-25%)	0	1	15	6
Low-medium (26-50%)	0	5	7	14
Medium-high (51-75%)	15	8	2	2
High (76-100%)	9	10	0	2

2.2 Students' reading behaviors on different item types

Drawing upon evidence from the eye-tracking, self-report checklist and interview data, we now present the results regarding students' reading processes on each item type one by one. We also refer to the GEPT Information Brochure and GEPT Level Descriptors (LTTC, 2016) to discuss the reading processes intended for each item type.

Before we present the eye tracking data, it is important to note that, as is the case in much eye tracking research, not all students' gaze behaviour on each test item was consistent or regular. In some cases, the technology could not capture a student's gaze movements fully on every test item, e.g. if they looked away from the screen frequently, or for other reasons their gaze behaviour was inconsistent. In such cases it is important for quality reasons to exclude such data from detailed gaze analysis so as not to distort the dataset. For this reason, the data reported for some items below are smaller numbers than the full cohort of 24 students.

Cloze – MCQ

The Cloze MCQ questions were set in the High-Intermediate level (corresponding to CEFR B2). The MCQ format has been one of the most popular item types since the 1960s with all TOEFL and ELTS items utilising this format. According to test information (LTTC, 2016), the intended type of reading was careful local reading, e.g. inferring lexical meaning and understanding of syntactic structures. In terms of level of reading processes, test takers were expected to understand meaning at the clause and sentence levels. As shown in Table 3, most students in this study did very

well on this section, with a mean accuracy of 74.40%. Table 5 shows the results of students' self-report regarding their reading processes on the Cloze MCQ items. We can see that most students, aligning to the intended construct, reported that they aimed to understand specific information at local level. However, all medium-high scoring students (i.e. those who scored between 51%-75% in this section reported that they mostly relied on test-taking strategies, e.g. matching options to the vocabulary in the text, to answer the questions. Only 6.67% of them believed they read the text carefully and 13.33% believed they focused at the sentence level. In other words, most the these medium-high students searched the text back and forth to identify the correct answers. In contrast, almost 80% of the high-scoring students believed that they mostly identified the answers by reading at the level of individual sentences/within sentences.

Table 5. Students' self-report checklist data on Cloze MCQ

		Low scoring (n=0)	Low-medium scoring (n=0)	Medium-high scoring students (n=15)	High scoring students (n=9)
Reading goals	To understand specific information*	N/A	N/A	100%	77.78%
	To understand main idea of the paragraph	N/A	N/A	13.33%	22.22%
Types of reading	Careful reading*	N/A	N/A	6.67%	55.56%
	Search reading	N/A	N/A	80%	44.44%
	Test-taking strategies	N/A	N/A	100%	22.22%
Levels of reading	Within sentence/sentence*	N/A	N/A	13.33%	77.78%
	Across sentences	N/A	N/A	93.33%	44.44%

Note: * indicates the intended reading processes for this item type

The interview data provided some insight into students' beliefs of the construct of this item type. It was clear that some high-achieving students saw the centrality of grammar knowledge in this section, with one of them, for example, remarking:

I focused on my grammar knowledge when deciding the answer and to check if the answer makes the sentence coherent.

Several of them noted that this section required 'integrating information across sentences', since it called also for inter-sentential analysis. As one student said:

I pay attention to the sentence before and after the blank.

However, some students seemed to misinterpret the construct of this item type. For example, one said,

I believe in this section you have to understand the meaning of every sentence, and the relation between sentences to fill in the blank.

While integrating information across sentences might help to answer the questions, the Cloze MCQ items in this section mainly targeted at careful local reading abilities such as syntactic parsing, as shown in the example (Q3) below.

Children's triathlons also differ ..... (Q3)..... they place less emphasis on competition than on participation.
A. much from
B. in that (correct answer)
C. with which
D. other than

To gain a more in-depth understanding of students' reading behavior, we examined successful and unsuccessful students' eye tracking data on Question 3. It was selected because it offered a clear focus on syntactic parsing and grammar knowledge, and was answered correctly by a relatively small number of students (8 correct and 16 incorrect).

Question 3 was analysed by the researchers as testing syntactic parsing for the most part. Students were expected to a) to read the whole sentence carefully, and then b) to

use their grammatical knowledge to distinguish between the correct and incorrect options We compared the successful students' gaze data with gaze data from unsuccessful students on three areas of interest, including the target sentence, the correct option and incorrect options. To ensure complete accuracy of analysis, students whose gaze date on this item was in any way unusual, defective or unclear were excluded. Therefore, data from a total of 4 successful and 13 unsuccessful students' gaze activity was used, see Table 6.

Table 6. Gaze data results for Question 3

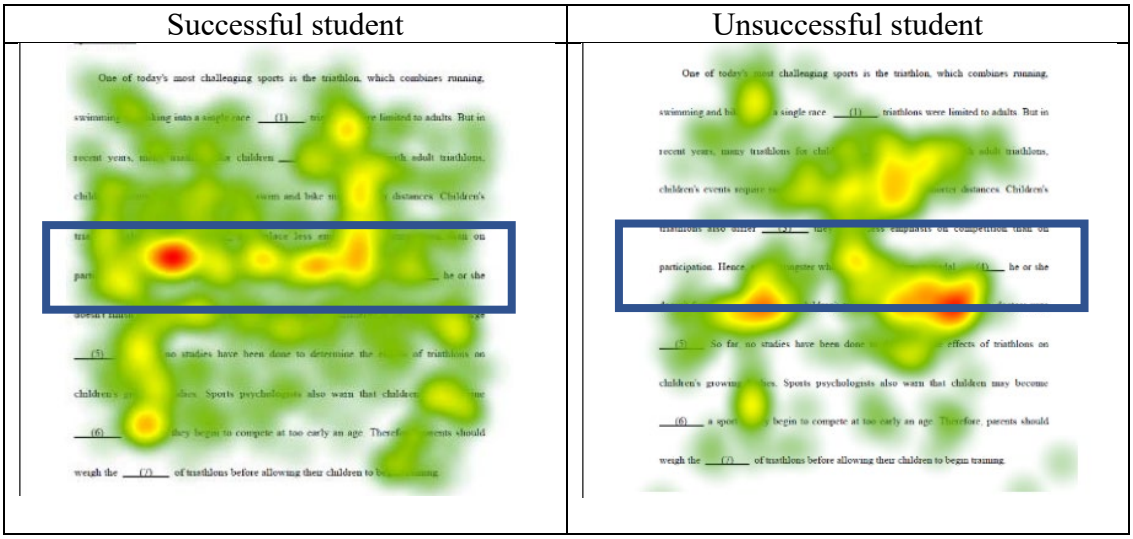
	Mean Fixation Duration (seconds)		Mean Fixation Count (number)		Mean Visit Duration (seconds)		Mean Visit Count (number)	
	S (n=4)	U (n=13)	S (n=4)	U (n=13)	S (n=4)	U (n=13)	S (n=4)	U (n=13)
Target sentence	M=13.80 SD=7.39	M=8.34 SD=4.61	M=76.50 SD=37.14	M=48.00 SD=25.56	M=1.10 SD=0.56	M=0.73 SD=0.35	M=19.50 SD=9.00	M=17.62 SD=10.10
Correct option	M=2.62 SD=0.82	M=2.47 SD=1.97	M=12.25 SD=4.92	M=12.00 SD=8.31	M=0.47 SD=0.27	M=0.44 SD=0.25	M=7.25 SD=2.22	M=6.67 SD=3.63
Incorrect options	M=7.75 SD=4.11	M=10.91 SD=6.00	M=32.05 SD=15.44	M=33.08 SD=16.12	M=1.08 SD=0.51	M=0.86 SD=0.34	M=17.75 SD=9.54	M=17.55 SD=8.57

Note: S= Successful students; U= Unsuccessful students

The findings showed a modest difference in terms of overall attention paid to the text, with the successful students appearing to fixate more extensively and frequently on the target text. Successful students also fixated more on the correct options than the unsuccessful students in terms of all the measures - fixation duration, fixation count, visit duration and visit count. As can be seen in Table 6, successful students outranked unsuccessful students on these measures, showing that they attended more closely to the target text and the correct option (i.e. indicating careful local reading as they reported). In contrast, unsuccessful students tended to fixate more intensively and visit more frequently the incorrect options which they were about to choose (i.e. indicating searching vocabulary in the text which might match options).

Qualitatively, Figure 1 offers an interesting illustration of the trend discussed above with Question 3, in that the successful student, whose gaze patterns are presented as the heatmap on the left, clearly focused far more intensively on all parts of the target sentence around the item (indicated by a rectangle) than did the sample unsuccessful student on the right, whose focus on that area was patchy.

Figure 1. Heatmap of gaze behavior of a successful and unsuccessful student on Q3



Note: the heat map indicates how students' fixations are distributed over the stimulus. Red indicates intensive fixations.

In summary, then, cloze MCQ items in this study intended to assess students' abilities in careful local reading, for example with Question 3 focusing on syntactic parsing and grammar knowledge. Although most students did well in this section as a whole, there was some evidence that successful students fixated on and visited the target text and the correct test option more than unsuccessful students, and were rewarded for this. Additionally, high-scoring students were inclined to employ the intended reading processes, i.e. careful local reading at sentence level, whereas the

medium-high scoring students tended to get the correct answers by unintended test-taking strategies.

Reading comprehension – MCQ

The reading comprehension MCQ questions were set in the High-Intermediate level (corresponding to CEFR B2) of the GEPT reading test. In this section, students were required to read a longer article and answer comprehension questions. According to test information (LTTC, 2016), this section intended to test students' careful reading skills at both local and global levels, tapping into a range of reading processes. As shown in Table 2 above, students scored highly on this section with a mean of 71.43%, and indeed students reported that they found it relatively straightforward. A good description of how most students approached this section was offered by a high-scoring student:

> *I got the overall meaning of the whole article by reading the first and last sentence of each paragraph quickly. I then read the questions and located the relevant parts in the article for each question. The article was structured in sequence so it wasn't too difficult to locate the relevant parts.*

This illustrates the element of search reading used in this part of the test, and the need for *establishing propositional meaning at clause and sentence level*, of *integrating information across sentences* and also of *creating a text level structure* for some of the test items.

The self-report checklist data (see Table 7) showed that the medium-high and high scoring students recognised the need to employ different reading skills in this section.

In contrast, most low-medium students focused on careful reading to understand main ideas of the whole text (which was the intended focus of the final item in this section - see the discussion of Q14 below). The low scoring student reported that he/she did not use careful reading but relied on search reading and test-taking strategies.

Table 7. Students' self-report checklist data on Comprehension MCQs

		Low scoring students (n=1)	Low-medium students (n=5)	Medium-high scoring students (n=8)	High scoring students (n=10)
Reading goals	To understand specific information*	0.00	0.00	62.50	100.00
	To understand main idea of paragraph*	100	20.00	100.00	100.00
	To understand main idea of whole text	0.00	100	25.00	0.00
Types of reading	Careful reading*	0.00	100.00	100.00	100.00
	Search reading *	100.00	20.00	75.00	100.00
	Test-taking strategies	100.00	40.00	12.50	0.00
Levels of reading	Within sentence/sentence*	100.00	0.00	0.00	0.00
	Across sentences*	100.00	100.00	100.00	100.00
	Text-level*	0.00	40.00	100.00	20.00

Note: * indicates the intended reading processes for this item type

For qualitative analysis, we identified two test items to analyse in detail, one which required close attention to a specific part of the text (Q13) and one which required understanding of the whole text (Q14). Investigating these items was therefore important in our attempt to understand the whole range of cognitive processes (see Table 1), since they required respectively "establishing propositional meaning at clause and sentence level" and "creating a text level structure". Question 13 was as follows:

What was van Meegeren originally accused of doing?

A. Assisting foreigners to obtain a national treasure (correct answer)
B. Trading fake paintings for special privileges
C. Stealing paintings done by Vermeer
D. Telling the police a series of lies

This requireed students to locate the relevant part of the passage by search reading, (identified as the 'target text'), namely:

> *Meegeren's fake Vermeer painting was sold to a German officer. After the war, van Meegeren was arrested by Dutch police for enabling this Dutch "masterpiece" to fall into German hands, a serious crime.*

They would then need to establish the propositional meaning of individual sentences and integrate information across the two sentences in the target text. To answer this question correctly, students would also need to use lexical knowledge to recognise the link between, for example, 'accused' in the question and 'arrested' in the text, and '*national treasure*' in the correct option and '*Dutch masterpiece* ' in the text.

In terms of eye gaze measures, when the Mann-Whitney U test was applied to test differences between successful and unsuccessful students' gaze patterns on the key areas of interest, no significant difference was noted between successful and unsuccessful students in terms of their attention to the target sentence or correct and incorrect options, with one exception. This was in terms of visit duration on the target text sentence that the successful students spent significantly longer on each visit to the target text than unsuccessful students (see Table 8).

Table 8. Visit duration on target text (Q13)

	N	Median	Mean	SD	Mann-Whitney U	Wilcoxon W	Z	p (2-tailed) p<0.05
Successful	10	1.05	1.21	0.66	8.00	23.00	-2.08	.037
Unsuccessful	5	0.50	0.55	0.18				

Once again caution should be used in drawing firm conclusions from this, given the small sample, but it would appear to support the conclusion drawn in Bax (2013) and Brunfaut and McCray (2015) that successful students are better able to find and identify the appropriate part of a text, using search reading, and that they can then read it careful, while unsuccessful students, who are less able to set the appropriate reading goal, fail to do so.

As noted previously, the second item we identified from this task (comprehension MCQs) required understanding of the whole text (Q14). This was an important item in our analysis, since it intended to test the higher-level reading processes (see Table 1) of "creating a text level representation".

> *What does the article indicate about van Meegeren?*
> *A. His reputation as an artist surpassed Vermeer's.*
> *B. He showed a painting by Vermeer in court.*
> *C. He was admired for his artistic talent.*
> *D. The charge against him was reduced. (correct answer)*

Clearly, although the correct answer was to be found in the last paragraph, the distractors oblige students to read the whole text so as to eliminate incorrect possibilities, as well as to use their lexical and syntactic knowledge. For this reason, the areas of interest selected for analysis with this item included the text as a whole, as well as the final paragraph, the target sentence itself, and the question (distinguishing between correct and incorrect answers). The key paragraph and target sentence (underlined here, but not in the original) were:

> *At his trial, van Meegeren confessed that the 'masterpiece' in question was a fake. No one, however, believed him. Van Meegeren finally convinced the*

judge by painting another fake Vermeer. Consequently, van Meegeren was convicted on a lesser offense - forging an artist's signature - and sentenced to a year in prison. His case fascinated the public and revealed how easily even experts can be deceived by fakes that bear famous names.

Results from the eye tracking gaze measures showed no significant differences between successful and unsuccessful students in terms of attention paid to the question itself. However, they did show a significant difference in terms of attention paid to one page of the text (the first one of two in the onscreen version), and also to the target paragraph on the second page, and (separately) to the target sentence. In all three cases, when the Mann-Whitney U test was applied to the data from students whose gaze data was of acceptable quality, significant differences were noted between successful and unsuccessful students. Successful students spent significantly longer on these three target areas than unsuccessful students (see Table 9). Specifically, successful students showed more fixations on page 1 of the text (fixation count mean - indicating intended reading), and on each visit which they made to both the target paragraph and sentence they spent significantly longer than unsuccessful students (visit duration mean).

Table 9. Statistics of gaze measures for Q14

Text page 1 *Fixation Count Mean*

	N	Median	Mean	SD	Mann-Whitney U	Wilcoxon W	Z	p (2-tailed) p<0.05
Successful	8	269.00	283.63	89.53	15.00	60.00	-2.02	0.43
Unsuccessful	9	197.00	207.11	80.07				

Target paragraph *Visit Duration Mean*

	N	Median	Mean	SD	Mann-Whitney U	Wilcoxon W	Z	p (2-tailed) p<0.05
Successful	8	8.13	6.62	2.79	11.0	56.0	-2.406	.016
Unsuccessful	9	2.69	3.16	2.15				

Target sentence *Visit Duration Mean*

	N	Median	Mean	SD	Mann-Whitney U	Wilcoxon W	Z	p (2-tailed) p<0.05
Successful	8	1.67	2.07	0.98	10.00	46.00	-2.312	.021
Unsuccessful	9	0.91	1.00	0.48				

It is not methodologically sound to attribute too great a significance to these results, given the small sample, but they are nonetheless indicative of the phenomenon that to answer this question, successful students tended to employ the appropriate reading processes, i.e. to create a text-level representation by careful reading (as compared to scanning for matching vocabulary and using other test-taking strategies as reported by the low and low-medium scoring students), and to identify and then pay notably greater attention to relevant parts of the text (as was also seen with Q13 above). This indicated that Question 14, which aimed to test higher-level reading processes in Khalifa and Weir's terms, was indeed successful in distinguishing between students who could create a text-level representation and identify key information from longer texts, and those who could not.

Careful reading: Summary Cloze

The summary task was set in the Advanced level (corresponding to CEFR C1) of the GEPT reading test. In this section, students were required to read a text and then complete a summary cloze passage. According to test information (LTTC, 2016), this section intended to test students' abilities to distinguish the main ideas from supporting details in a text and to transform that text-level representation by completing a summary cloze. It will be recalled that student performance on this

section was the weakest. Almost all students scored less than 50% in this section, with 15 of them scored less than 25%.

The self-report checklist data clearly indicate that the students (both the low and low-medium groups) reported that they attempted to use all sorts of reading processes, but obviously unsuccessfully, in this section. At interview, the participants noted that the summary task was particularly difficult (see Section 4). Although summarisation is a natural entailment of reading comprehension (Kintsch & van Dijk 1978, van Dijk & Kintsch 1983), it does not necessarily happen automatically (Johns 1983). Summarization requires students to process beyond the level of comprehension with the additional need to read for evaluation, condensation, and transformation of ideas that have been presented (Hidi & Anderson 1986, p.473-74).

Table 10. Students' self-report checklist data on Summary

		Low scoring students (n=15)	Low-medium students (n=7)	Medium-high scoring students (n=2)	High scoring students (n=0)
Reading goals	To understand specific information	26.67	0.00	0.00	N/A
	To understand main idea of paragraph*	93.33	71.43	100.00	N/A
	To understand main idea of whole text*	40.00	100.00	100.00	N/A
Types of reading	Careful reading*	53.33	100.00	100.00	N/A
	Search reading	40.00	14.29	100.00	N/A
	Test-taking strategies	86.67	28.57	0.00	N/A
	Inferencing*	13.33	71.43	0.00	N/A
Levels of reading	Within sentence/sentence	20.00	0.00	0.00	N/A
	Across sentences	80.00	71.43	100.00	N/A
	Text-level	80.00	100.00	100.00	N/A

Note: * indicates the intended reading processes for this item type

The summary cloze here did not merely make use of words copied from the original text; as one student noted:

> *The summary has been paraphrased and most of the words have been changed from the original passage.*

This added a degree of complexity to connect their own text-level representation of the original text and the summary cloze, with an additional lexical and syntactic parsing load. While some students recognised the need to connect the two, they were not able to do so, as one student recalled:

> *I was then trying to identify some key words in the summary which link back to the article. I was lost.*

One student tried to use not only lexical but meta-discourse devices to assist:

> *To find the answer, I tried to focus on the connectives, e.g. however, to guide me to the relevant part. But it was difficult, I couldn't find most of the answers.*

Indeed, many weaker students reported that they fell back on test-taking strategies by trying to match words in the summary and the passage. For example, one low-scoring student reported as follows:

> *I tried to match some key words between the summary and the original text, e.g. authority and the UK government.*

It was clear from these examples that such lower level reading strategies focusing at the lexical level were proved to be inadequate with this item type. By contrast, more successful students (medium-high scoring group) specifically reported using higher-level reading processes such as inferencing to fill the gaps between the two representations (i.e. their own representation of the original text and the summary cloze), as in the example below:

> *Sometimes I made a guess based on my understanding. For example, it was talking about industries closing down so it means they moved out of the area.*

For the eye-tracking data analysis, Question 19 was identified as it was a good representative of the items as a whole. The relevant section of which for this item was this:

> *Today, success stories can be found in many urban areas,*
> *where .....(19)....., commercial spaces, and recreational facilities now stand on land that once contained only deserted buildings and parking areas*

The answers permitted by the mark scheme included "residences, housing, homes, housing estates, the construction of new housing/homes/houses". In order to identify any of them students would need as a minimum a) to find the relevant parts of the text, and (b) to use their knowledge of lexis, and in particular of synonymy, to match the gap in the cloze text and at the same time to eliminate a number of important distractors (e.g. 'commercial spaces and recreational facilities'). In short, this was a challenging item, drawing on a number of high-level cognitive processes, including word recognition, lexical access, integrating information across sentences and inferencing (see Table 1).

The eye tracking gaze data did not reveal significant differences between the successful and unsuccessful students on this item in terms of focus on or visits to the cloze text itself, but it did reveal a significant difference in terms of one measure, namely visit count, in relation to the specific target phrasing (i.e. the precise short passage in the text itself which students had to identify and then read intensively). The statistics can be found in Table 11.

Table 11. Visit Count on Target Phrasing on Q19

	N	Median	Mean	SD	Mann-Whitney U	Wilcoxon W	Z	p (2-tailed) p<0.05
Successful	5	12.00	10.90	4.18	11.00	89.00	-2.01	0.44
Unsuccessful	12	6.50	6.25	5.34				

This implied that successful students had identified that this was a key area to which they had to attend, and then came back to it repeatedly, significantly more than unsuccessful students did. This is illustrated in Figure 2, where it can be seen that one successful student had paid far more attention to the target phrasing than the unsuccessful student.

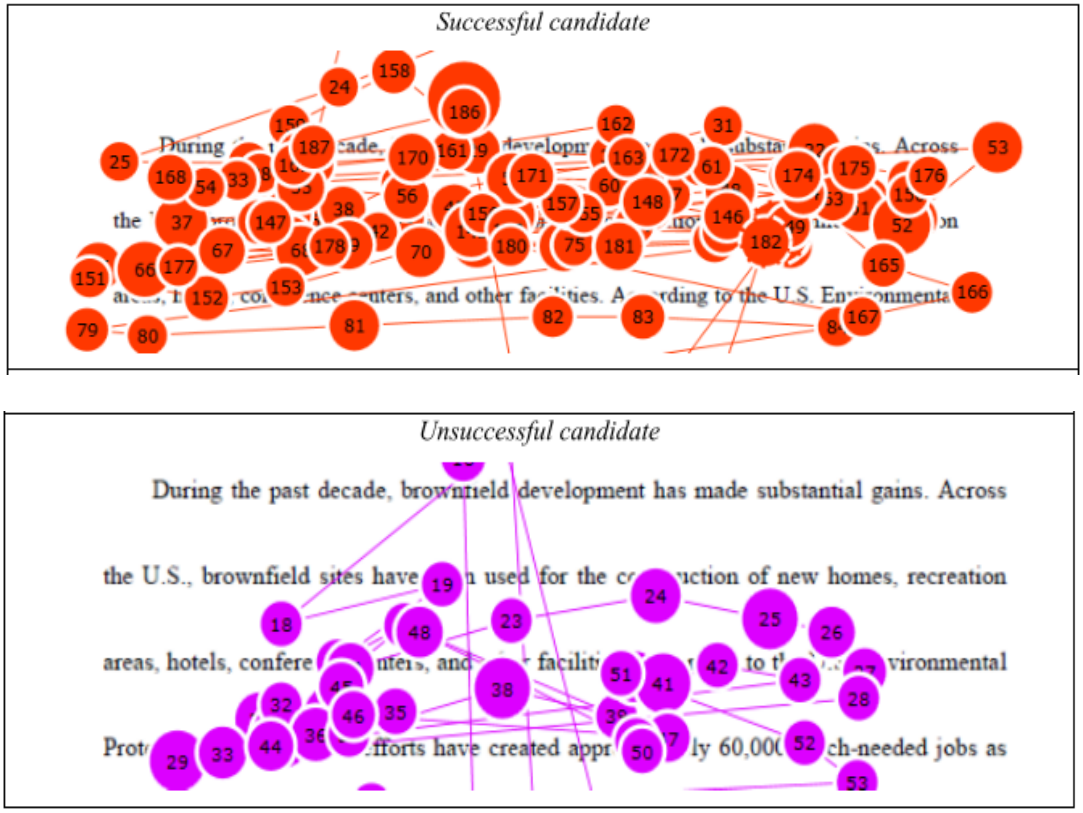


Figure 2. Gaze data for Target Text for Q19

Note: The diameter of the circles indicates the fixation duration. The longer the fixation, the larger the circle. The number on the circles indicate the time sequence of fixation.

As mentioned previously, this task type (summary) required students to create a text-level representation of the original text by identifying the main ideas, and to connect their representation with the summary cloze. Therefore, we would expect the students to read the cloze section and then return to read the target section of the text intently a number of times. This data showed that, with this item, the successful students did precisely that, although with these small samples such an analysis must remain suggestive rather than conclusive.

In summary, although the summary cloze task led to some frustration among students, and a sense of pressure, from a test design point of view the data demonstrate a high degree of effectiveness on the part of this set of test items, since they clearly not only required higher-level cognitive skills, but then rewarded those who used them with better marks.

Expeditious reading

The final item types, also at Advanced level, specifically targeted students' expeditious reading skills, i.e. skimming and scanning. As mentioned earlier, GEPT is one of the few high-stakes reading examinations which dedicate items to assessing this important aspect of reading skills. The self-report checklist data (see Table 12) showed that students at all levels of performance reported carrying out the kinds of reading processes expected of these items (i.e. search reading at textual and intertextual levels). However, the low and low-medium groups reported using careful reading too. The direct consequence of using careful reading in this section was that these students could not finish all questions within the time allowance.

Table 12. Students' self-report checklist data on Expeditious Reading items

		Low scoring students (n=6)	Low-medium students (n=14)	Medium-high scoring students (n=2)	High scoring students (n=2)
Reading goals	To understand specific information	66.67	71.43	0.00	0.00
	To understand main idea of paragraph*	66.67	71.43	0.00	100.00
	To understand main idea of whole text*	16.67	0.00	100.00	100.00
Types of reading	Careful reading	50.00	50.00	100.00	0.00
	Search reading *	83.33	100.00	100.00	100.00
	Test-taking strategies	67.67	28.57	0.00	0.00
	Inferencing*	0.00	0.00	0.00	100.00
Levels of reading	Within sentence/sentence	16.67	7.14	100.00	0.00
	Across sentences	66.67	92.86	100.00	50.00
	Text-level*	66.67	50.00	100.00	100.00
	Multi-texts level*	16.67	7.14	0.00	100.00

Note: * indicates the intended reading processes for this item type

The first five questions in this section requireed students to choose the appropriate heading from a bank of possible headings for each paragraph of the text. The stimulated recall data showed that students scanned and skimmed the texts at varying degrees of sophistication. To start with, this lower scoring student reported a relatively basic cognitive operation:

I just skim the paragraph and choose my answer based on some keywords I noticed.

By contrast, this higher scoring student reported rather more sophistication and detail:

I first skimmed the headings to identify some keywords to get me a general idea and then checked if I could see these keywords or similar words in every paragraph. I went back and forth to see if I could match the keywords.

In a task of this kind, however, simple reliance on the low-level cognitive process of matching a few key words could be risky. One high scoring student recognised and reported this explicitly, and at the same time demonstrated a remarkable awareness of her use of higher-level reading processes to complete the test items successfully:

> *I needed to understand the overall meaning of each paragraph before choosing the most suitable heading. Some headings were quite similar so I had to understand the meaning of the paragraph. If I only relied on some key words, they would probably misguide me. (emphasis added)*

Furthermore, it was clear from this report that the student also made use of high-level inter-sentential reading, as well as search reading

For eye-tracking data, Question 21 and Question were selected. Question 21 required students to read a set of paragraphs and then to match them with the correct heading from a set of 11, a cognitively demanding task requiring at least word recognition, lexical access, integrating information across sentences and inferencing (see Table 1). The five areas of interest which were analysed consisted of the whole text paragraph, two key items of lexis, the set of answers as a whole, and the correct answer.

There was no significant difference identified between successful and unsuccessful students on most of these areas, with the exception of visit duration in the relevant paragraph itself, in other words the whole target paragraph which readers had to read so as to identify the correct heading. Successful students spent significantly longer in each visit they made than unsuccessful students, as can be seen in Table 12.

Table 12. Visit Duration on Target Paragraph on Q21

	N	Median	Mean	SD	Mann-Whitney U	Wilcoxon W	Z	p (2-tailed) p<0.05
Successful	8	6.78	7.31	2.56	12.00	48.00	-2.10	0.36
Unsuccessful	8	4.26	5.25	3.75				

This indicated that successful students were obliged by the test item to pay close attention to the relevant paragraph, and did so productively, in comparison with unsuccessful students. This in turn suggested that the test item was working effectively and achieving its aims.

On the other hand, Question 30 required students to read a set of longer texts and then to answer a question which required understanding of detail in the text. In this case the question was:

Which historical attraction offers theatrical productions?

The answer was in the first of the three texts, which included the lexis 'street performances', 'dramatic events' and 'acted out' to match the term 'theatrical productions' in the test question.

To answer correctly would require the students to read all three texts and use at least the cognitive processes identified by Khalifa and Weir (2009) as word recognition, lexical access, integrating information across sentences and inferencing. In addition, they would need to some extent to use the most complex of processes, namely integrating information across texts (see Table 1) since they would need to contrast the information in each of the longer texts in order to rule out possible wrong answers. Four areas of interest were analysed, focussing respectively on the set of

questions, the specific Test question for Question 30, one key phrase in the text with key pieces of lexis, and the whole of the correct page of text.

In the event the analysis identified no significant differences between successful and unsuccessful students in terms of their reading of the test questions or the target phrase with key lexis, but it did identify a significant difference in Visit Count for the relevant page on which the answer was to be found (see Table 13).

Table 13. Visit Count on the correct text on Q30

	N	Median	Mean	SD	Mann-Whitney U	Wilcoxon W	Z	p (2-tailed) p<0.05
Successful	9	12.00	13.78	7.07	12.50	40.50	-2.03	.043
Unsuccessful	7	6.00	8.00	2.67				

This means that successful students made significantly more visits to the relevant page of text than unsuccessful students. It is not possible to say this was the result of their work on Question 30 alone, since they were also reading for the other items in that section, and the dataset was too small to allow a firm conclusion, but it nonetheless implied that the successful readers were more active in their reading in terms of number of visits to this page. However, no statistically significant differences were found in terms of the time they actually spent on that page.

5. Discussions

Was eye-tracking useful in revealing students' reading processes?

The data gathered from the eye tracking technology was illuminating in several ways. It must be reiterated once again that the samples are small, owing to the fact that data from each student has to be gathered individually, in a highly time-consuming

process, and for this reason the research findings must be seen as indicative as opposed to conclusive.

Although each reading item analysed above produced different measures, much of the data pointed in the same direction, namely that successful students on each test item were probably successful because they focussed their gaze either *for longer*, or *more frequently*, or *more productively* on key areas of the test items or the texts. This in turn suggests that they were better able to identify those areas on which it is most strategic to focus. An important caveat when dealing with eye tracking data is the fact that it is not always possible with confidence to interpret readers' cognitive processes using gaze data alone, for which reason it was important in this project also to collect self-reports and stimulated recall data in the way described in the methodology section above. Nonetheless, the eye-tracking data was useful in indicating – in line with previous literature – that the gaze behaviour of successful L2 students in test conditions does differ significantly at key points from the gaze behaviour of unsuccessful students (Bax, 2013; Brunfaut & McCray 2015), in ways which can help us to understand how better to train readers and how better to test them.

Did different item types elicit different reading processes?

A further important finding from the eye tracking data is that the GEPT test appears to be functioning effectively in terms of cognitive processing in the areas analysed. It is clear from all of the reading item types discussed above that they were

successfully leading students to carry out a range of lower-level reading processes (e.g. Question 3) and also reading activities (e.g. Questions 13, 14, 19, 21 and 30) requiring the type of high level cognitive processing identified by Khalifa and Weir (2009) and others as appropriate for testing reading at higher intermediate and advanced levels.

The findings indicate that item types have an impact on students' reading processes and use of test-taking strategies, indicating the importance of selecting appropriate item types to test particular reading skills. It is possible also conclude that the Advanced sections of the test elicited the same set of cognitive processes as the High-Intermediate test, with the addition in the final section of the most difficult of all in Khalifa and Weir's (2009) scheme, namely integrating information across texts. However, this last element was only tested partially, since students in that section needed to read across different texts only so as to exclude possible wrong answers, and not precisely to "integrate information" in any complex way. If the test designers wished to raise the cognitive level of this Advanced test, then a more fully 'intertextual' activity could arguably be added to the GEPT Advanced paper, by which students could be asked to demonstrate that they had read and assimilated information across more than one complex text.

It is also important to note that in each case it was the successful test takers who exhibited the target cognitive processing, meaning that the test items are effective in distinguishing them from those test takers who do not employ the relevant cognitive

processes. This is an important new piece of evidence in building a validity argument for the GEPT tests under investigation.

Nonetheless, an issue arose from the findings was students' use of test-taking strategies, some of which were arguably construct irrelevant. If we consider the implications of the self-report and stimulated recall interview data, it is clear from the information in Tables 5, 7, 10 and 12 above, and also from interview data presented above that lower-scoring participants tended to report the use of more test-taking strategies, e.g. matching key words, focusing on topic sentences, getting hints from grammar, eliminating options, etc. This can be seen for example in the Cloze and Comprehension MCQ sections, where 100% of the lower scoring students reported using such strategies.

When asked about test-taking strategies in the retrospective interviews, students reported that they were aware that these were often not the best strategies to use, especially on more advanced tasks, and admitted that the strategy they used might not be effective for a particular section. It appeared as if they were aware of the kinds of reading which the more difficult items required, but that they were not always able to employ the appropriate reading cognitive processes, and instead fell back on pre-taught test-taking strategies. This occurred to the extent that some of the low-scoring students reported using the same test-taking strategies for all parts of the test, whilst realising that this was unlikely to be productive. One possible reason for this, alluded to by some participants, was a negative washback effect resulting from their

extensive test preparation (for all tests in general rather than specifically for GEPT) at school over the years.

This might suggest that, for item design in future, the test designers could usefully revisit the use of MCQs, since it appeared from the interview data that students used test-taking strategies disproportionally on MCQs, hoping to achieve success not by reading the text carefully but by working out the correct option from concentrating extensively on the MCQ options given to them.

6. Conclusions

The purpose of this study was to examine the relationship between reading item types and students' reading processes in test conditions. In isolation, the eye-tracing data cannot positively identify the cognitive processes employed by the reader from one moment to the next. However, eye-tracking data, when corroborated with stimulated recall interview and questionnaire data, does give a strong indication of the cognitive processes employed. It is legitimate to infer that if a test item requires a certain cognitive process from a student (e.g. inferencing), and the student answers the item correctly, then also offers gaze data to show the appropriate eye movements (e.g. focusing on the target sentence), and then also reports having read the text in that way, we can reasonably infer that the target cognitive process has been used.

Building on Bax's (2013) seminal study on eye-tracking and test taking, this study adds to the compelling evidence of the value of using eye tracking in conjunction

with other methods for researching cognitive validity in reading tests. Additionally, this study further provides empirical evidence illumining the impact of common reading item types on students' reading processes. In line with previous studies on reading tests (Enright et al 2000; Owen, 2016; Weir et al, 2000), this gives a clear indication of the need for a reading test to include various reading item types, especially those targeting at global expeditious reading and creating representation at the intertextual level.

Although this study yielded some interesting insights into the relationships between reading item types and students' reading processes, it has a number of limitations. First, the background of the participants was homogeneous, both in terms of nationality and level of reading proficiency. This might have impacted on the possible variation of the participants' reading processes on the item types and thus limited the genenalisibility of the findings. A potential avenue for future research would be replicating the research with participants from a wider range of L1 backgrounds and levels of reading proficiency. A cross-linguistic approach might provide insight into how orthographic characteristics of different languages influence the impact of item types on students' reading processes.

Second, only five item types (i.e. cloze, reading comprehension multiple chose questions, summary, heading matching and which text?) from one version of the GEPT High Intermediate and Academic reading tests were used to elicit reading

performances. An important area of follow-up research would be to repeat the study utilising other reading item types.

Finally, as the purpose of the research was to inform language test validity and improve item design, our analysis was largely guided by socio-cognitive model of reading. However, a closer link between computational models of eye movement (for example see Reichle, Rayner, & Pollatsek, 2003) and reading models for language test development could be forged to extend our understanding of students' reading processes at different proficiency levels in future studies .

References

AUTHOR (DATE)

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal, 91*(4), 659-663.

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

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

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. *Language Testing, 24*(2), 209-50.

Cooper, M. & Holzman, M. (1983). Talking about protocols. *College Composition and Communication, 34*(3), 284-293.

Eger, N., L. Ball, R. Stevens & Dodd, J. (2007). *Cueing retrospective verbal reports in usability testing through eye-movement replay.* Proceedings of HCI 2007, The 21st British HCI Group Annual Conference University of Lancaster, UK.

Enright, M, Grabe, W, Koda, K, Mosenthal, P, Mulcany-Ernt, P and Schedl, M (2000) *TOEFL 2000 Reading Framework: A Working Paper*, TOEFL Monograph Series 17, Princeton, NJ: ETS.

Glaser, R. (1991). *Expertise and assessment*. In M. C. Wittrock and E. L. Baker (Eds), *Testing and cognition* (pp. 17-30)Prentice Hall, Englewood Cliffs.

Grabe, W (1991) Current developments in second language reading research, *TESOL Quarterly* 25 (3), 375-406.

Grabe, W., & Stoller, F. L. (2002). *Teaching and Researching Reading*. Harlow: Pearson Education.

Hidi, S., & Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of Educational Research*, 56, 473–493.

Hinton, P. (1995). *Statistics Explained: A Guide for Social Science Students*. London: Routledge Psychology Press.

Johns, A. M. (1993). *Reading and writing tasks in English for academic purposes classes: Products, processes and resources*. In J. G. Carson & I. Leki (Eds.), Reading in the composition classroom (pp. 274–289). New York: Newbury House.

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

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363–394.

LTTC (2016) *GEPT Information Brochure*. Retrieved in May 2018 https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/files/GEPT_Information_Brochure.pdf

LTTC (2016) *Level descriptors.* Retrieved in May 2018 https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/hi_intermediate.htm

Owen, N. (2016). *An evidence-centred approach to Reverse Engineering: Comparative analysis of IELTS and TOEFL iBT reading sections.* Unpublished PhD thesis, University of Leicester.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, *26*, 445–526.

Ruiz-Primo, M., R. Shavelson, M. Li & Schultz, S. (2001). On the validity of cognitive interpretations of scores from alternative concept-mapping techniques. *Educational Assessment,* 7(2), 99-141.

Russo, J. E., E. J. Johnson, & Stephens, D. L. (1989). The validity of verbal protocol. *Memory and Cognition, 17*, 759-769.

Sheskin, D. (2003). *Handbook of Parametric and Nonparametric Statistical Procedures.* Boca Raton, Florida: CRC Press.

Tobii Technology AB (2013). *Product Description Version 1.0.1*, Retrieved from http://www.tobii.com/Global/Analysis/Downloads/Product_Descriptions/Tobii_X2_Pr oduct_Description.pdf?epslanguage=en

Van Dijk, T. A., & Kintsch, W. (1983). The notion of macrostructure. T. A. van Dijk & W. Kintsch (Eds.), *Strategies of discourse comprehension*, 189-223. New York: Academic Press.

Weir, C J, Yang, H and Jin, Y (2000) *An Empirical Investigation of the Componentiality of L2 Reading in English for Academic Purposes*, Studies in Language Testing 12, Cambridge: UCLES/Cambridge University Press

Weir, C. (2005). L*anguage Testing and Validation: An evidence-based approach.* London: Palgrave Macmillan.

Wood, A., P. Fletcher & Hughes, A. (1986). *Statistics in Language Studies.* Cambridge, England: Cambridge University Press.

Wu, R. Y. F. (2014). *Validating second language reading examinations: Establishing the validity of the GEPT through alignment with the Common European Framework of Reference.* Studies in Language Testing 41. Cambridge, England: Cambridge University Press.