

LOOKING INTO TEST-TAKERS' COGNITIVE PROCESSES WHILE COMPLETING READING TASKS:

A mixed-method eye-tracking and
stimulated recall study

AR-G/2015/001

Tineke Brunfaut and Gareth McCray
Lancaster University

ABSTRACT

This research examined the cognitive processing of 25 test-takers while completing Aptis reading tasks. The study investigated test-takers' task processing in general, and according to a number of task and test-taker characteristics.

More specifically, sub-analyses were conducted to explore potential differences in cognitive processing between tasks targeting different CEFR levels, and between test-takers of different L2 reading proficiency and overall L2 proficiency. To this end, a combination of eye-tracking and retrospective interviews with eye-tracking traces as stimuli was used. Test-takers' L2 (reading) proficiency was measured by means of the full Aptis test.

It was found that test-takers engaged in a wide range of cognitive processes while completing the Aptis reading tasks, including the lower- and higher-level processes defined in Khalifa and Weir (2009) (with the exception of intertextual representation). Although successful item completion was most often associated with a careful global and/or local reading approach, expeditious reading was conducted by some test-takers on some tasks. Only a few potential threats to the test's construct validity were identified, and these risks were associated with specific individual items, not with the tasks or test as a whole (so these may be solved at the item writing level).

Different patterns were observed in the main forms of processing used to complete the different CEFR-linked tasks, which seem to be largely related to task type (more so than CEFR target level). Although these patterns did not constitute threats to the overall Aptis component's construct validity, the B1 gap-fill tasks may at least partly elicit different cognitive reading processes than those set out to be tested with this specific task type. Some trends were also noticed in the processing conducted by test-takers of different levels of L2 (reading) proficiency, although these were weaker (potentially due to the sample of participants).

Overall, the data indicate that the Aptis reading component as a whole samples widely from the construct of reading. These findings provide key information for Aptis validation purposes.

Methodologically, the combined use of eye-tracking and stimulated recalls proved achievable and, moreover, fruitful. The two methods allowed balancing the strengths and weaknesses of each individual method, generating a richer and wider-reaching set of data than each alone, and allowing triangulation of the findings of each method.

Authors

Tineke Brunfaut

Dr Tineke Brunfaut lectures in language testing at Lancaster University, UK. Her main research interests are language testing, and reading and listening in a second or foreign language. She has conducted research on factors affecting academic reading proficiency and second language listening task difficulty, the use of eye-tracking to look into second language reading, diagnostic assessment, and standard setting. Her work has been published in journals such as *Applied Linguistics*, *Studies in Second Language Acquisition*, *TESOL Quarterly* and *Language Assessment Quarterly*. Dr Brunfaut has also been involved in test development in a range of languages and countries around the world.

Gareth McCray

Dr Gareth McCray's research interests are psychometric modelling and investigating reading through the use of eye-tracking. He is a Research Associate in the Statistics Department at Lancaster University, UK, where he is investigating the modelling of child development trajectories in developing countries, and creating selection algorithms to assemble items into tests based on concurrent measures of the intended construct. In his PhD research, Dr McCray looked into the statistical modelling of cognitive processing in reading in the context of language testing. He has also conducted several studies on modelling eye-tracking data.

CONTENTS

1. BACKGROUND	5
2. INVESTIGATING READING	5
2.1. Cognitive processing model of reading	5
2.2. Reading research methodology	8
3. RESEARCH QUESTIONS	11
4. METHODOLOGY	11
4.1. Participants	11
4.2. Materials	12
4.2.1. Reading tasks	12
4.2.2. Full Aptis test	14
4.3. Data collection methodology and procedures	15
4.4. Ethical procedures and consent	17
4.5. Data analyses	17
4.5.1. Eye-tracking analyses	17
4.5.2. Stimulated recall analyses	22
5. FINDINGS	24
5.1. Descriptive statistics	24
5.2. Eye-tracking	25
5.2.1. Eye-tracking findings on cognitive processes when completing items targeting different CEFR levels (RQ1a)	30
5.2.2. Eye-tracking findings on cognitive processes depending on test-takers' L2 (reading) proficiency (RQ1b & RQ1c)	34
5.3. Stimulated recall	37
5.3.1. Stimulated recall findings on cognitive processes during Aptis reading test completion (RQ1)	37
5.3.2. Stimulated recall findings on cognitive processes when completing items targeting different CEFR levels (RQ1a)	39
5.3.3. Stimulated recall findings on cognitive processes depending on test-takers' L2 (reading) proficiency (RQ1b & RQ1c)	43
6. DISCUSSION	48
7. CONCLUSION	51
BIBLIOGRAPHY	52

LIST OF FIGURES

Figure 1: Khalifa and Weir's model of cognitive processing in reading – adapted from Khalifa and Weir (2009, p. 43)	7
Figure 2: Key eye-tracking measures	10
Figure 3: Adapted Aptis reading task layouts	12
Figure 4: Flowchart of the first data collection session	15
Figure 5: Heat maps of Aptis reading tasks	26
Figure 6: Aptis reading test overview for candidates (British Council, 2013, p. 11)	49

LIST OF TABLES

Table 1: Aptis reading component structure	12
Table 2: Eye-tracking metrics	18
Table 3: Eye-tracking measure hypotheses in relation to reading task CEFR level (RQ1a)	20
Table 4: Eye-tracking measure hypotheses in relation to L2 reading proficiency (RQ1b) and overall L2 proficiency (RQ1c)	21
Table 5: Coding framework stimulated recalls	22
Table 6: Descriptive statistics – Aptis reading tasks used for eye-tracking and stimulated recall (n=25)	24
Table 7: Descriptive statistics – full Aptis system (n=25)	24
Table 8: Aptis components – CEFR levels of participants (n=25)	25
Table 9: Descriptive statistics – Eye-tracking measures	30
Table 10: Results eye-tracking analyses in relation to reading task CEFR level (RQ1a)	31
Table 11: Eye-tracking support for RQ1a hypotheses	32
Table 12: Results eye-tracker analyses in relation to L2 reading proficiency (RQ1b) and overall L2 proficiency(RQ1c)	34
Table 13: Eye-tracking support for RQ1b and RQ1c hypotheses	36
Table 14: Stimulated recall results on cognitive processes during Aptis reading test completion (RQ1)	37
Table 15: Stimulated recall results on cognitive processes when correctly completing items targeting different CEFR levels (RQ1a)	39
Table 16: Stimulated recall results on cognitive processes of correct items depending on test-takers' L2 reading proficiency (RQ1b)	44
Table 17: Stimulated recall results on cognitive processes of correct items depending on test-takers' L2 proficiency (RQ1c)	47

1. BACKGROUND

In August 2012, the British Council launched a new computer-based, modular English language testing system for adults (16+), called *Aptis*, aiming “to help organisations and institutions identify standards of English and select the staff or students with the right skills” (British Council, 2014).

The test combines a core grammar and vocabulary component with one or more skills (speaking, listening, writing, and reading), “[t]esting English levels from A1-C on the Common European Framework of Reference for Languages (CEFR)” (British Council, 2014). Underpinning the test’s development is O’Sullivan (2011) and O’Sullivan and Weir’s (2011) reconceptualization of Weir’s (2005) socio-cognitive framework for language test validation (O’Sullivan, 2012). In addition, a monitoring and evaluation programme has been set up, consisting of in-house and externally-funded research, to examine the validity of the *Aptis* test system (O’Sullivan, 2012).

The research reported on in this document was an externally-funded study under the *Aptis* Assessment Research Grants 2012 programme. It focuses on the reading component of the *Aptis* test system. More specifically, the project aimed to examine test-takers’ cognitive processing while responding to *Aptis* reading comprehension items.

2. INVESTIGATING READING

2.1. Cognitive processing model of reading

In the socio-cognitive approach to test validation (O’Sullivan & Weir, 2011), a central role is assigned to the test-taker with his/her individual and cognitive characteristics. Performance and subsequent inference of ability is recognised as resulting from an interaction between the test-taker and test task. With reference to the skill of reading, the approach has formed the basis for Khalifa and Weir’s (2009) model of cognitive processing in reading, which integrates cognitive and metacognitive processes with language and general knowledge into a heuristic for reading comprehension.

As can be seen in Figure 1 on page 7, Khalifa and Weir’s (2009) model has three main components – metacognitive activity, the central processing core, and the knowledge base – which contain various sub-processes. Metacognitive activity, as described by Khalifa and Weir (2009) involves setting goals, monitoring, and remediating where necessary. When setting goals, the reader decides upon the type(s) of reading needed to complete a specific task: *local* reading at the sentence and clause level, or more *global* reading to understand the text beyond sentence and clause level; and *careful* reading to comprehend all the information in a text to extract a complete meaning, or *expeditious* reading employing selective and efficient strategies to access only specific information required from the text. While reading, readers monitor that their reading is progressing in line with the generated goals, and breakdowns trigger remediation of reading behaviour where necessary.

The central processing core, represented in the middle column in Figure 1, comprises a hierarchical system of eight distinct cognitive processes that are thought to work together to result in reading comprehension:

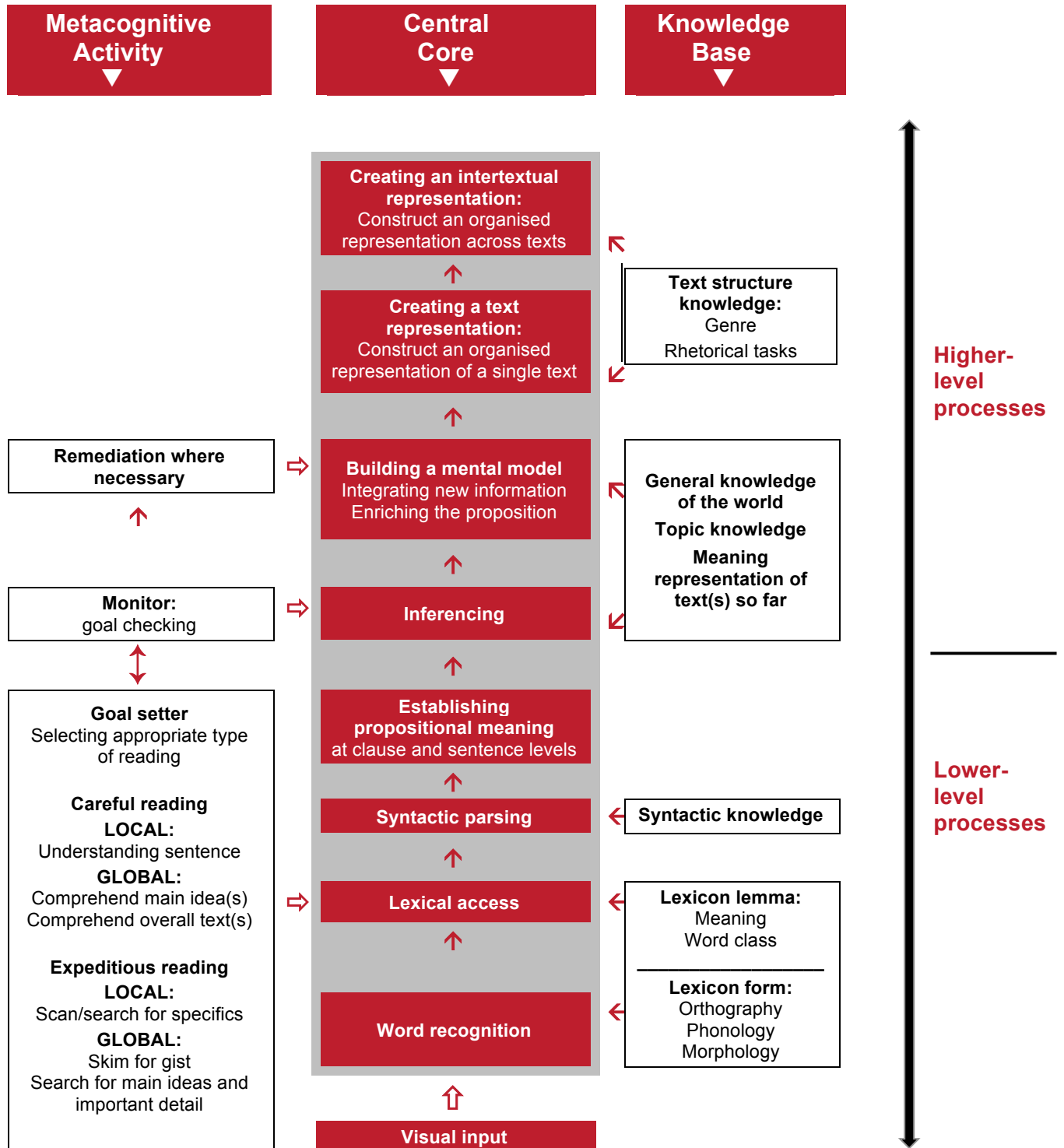
- the so-called lower-level processes – word recognition, lexical access, syntactic parsing, and establishing propositional meaning; and
- the higher level processes – inferencing, building a mental model, and creating a text level or intertextual representation (Khalifa & Weir, 2009).

The distinction between higher level and lower level skills, according to Grabe (2009), is the fact that lower level skills can become strongly automatised and not subject to conscious processing. Word recognition involves recognising the printed symbols (orthographic processing), sounding out the words in mind (phonological processing), and making use of information on expected grammatical forms (morphological processing). Lexical access concerns retrieval of information about the form and meaning of a word from the vocabulary stored in the reader's mind (the mental lexicon) to establish a word's meaning. Syntactic parsing involves the integration of the word at the clausal level, while deciphering the grammatical information in the text. In parallel with these processes, the clauses and sentences are converted into units of meaning to establish propositional meaning. Readers may also bring in their own knowledge of the world, of the topic of the text, and of the text itself to bear on the comprehension of the text. This process is referred to as inferencing. The integration of individual propositions into the overall meaning framework of the text is called the building of a mental model. This may lead to the creation of a text level representation whereby the text is constructed as a hierarchy of propositions, allowing the differentiation of the main points and gist of the text from its less significant details. Finally, information from multiple texts sources may be combined to create an intertextual representation.

While processing the text, the reader is likely to have various knowledge sources at their disposal, as depicted in the knowledge base in Figure 1, which link to specific aspects of the central processing core. These include: knowledge of a word's orthography, phonology and morphology; lexical knowledge of the meaning and the word class of a word; syntactic knowledge of the language; general knowledge of the world, topic knowledge about the subject of the text being read and knowledge of the text so far; and text structure knowledge, i.e. knowledge of genre or rhetorical forms.

The summarised model in Figure 1 (Khalifa & Weir, 2009) forms the theoretical foundation for the present study. The cognitive processing approach to reading represented in this model aligns with, and has been developed within, the socio-cognitive framework view for test validation (which, as mentioned above, has formed the basis of the Aptis development project). In addition, the Aptis reading tasks are developed according to a set of reading specifications which stipulate each task type's intended cognitive goal setting and processing level targets, as defined by Khalifa and Weir (2009) (for sample task specifications, see Dunlea, 2014, pp. 9–10). Therefore, in this study, inferences have been made on the empirical data of test-takers' reading processes during Aptis test completion in order to map these onto Khalifa and Weir's (2009) central processing core, while taking account of metacognitive, linguistic and general knowledge. This has not only allowed us to gain insights into test-takers' cognitive processing while responding to Aptis reading comprehension items *per se*, but also to evaluate the validity of a set of Aptis reading items as measures of the test's intended construct.

Figure 1: Khalifa and Weir's model of cognitive processing in reading – adapted from Khalifa and Weir (2009, p. 43)



2.2 Reading research methodology

In the socio-cognitive framework, Weir (2005) and O'Sullivan and Weir (2011) suggest that there may be a significant problem with a number of existing test validation procedures in that they only provide theory-based evidence for validity, via various statistical methods, based on test administrations. Qualitative insights are often restricted to "what the test constructors believe an item to be testing" (Alderson, 2000, p. 97), despite the many limitations and issues that have been pointed out regarding the use of expert judgement insights (e.g. Alderson, 1993; Alderson, Brunfaut, McCray & Nieminen, 2012). In the case of testing reading, researchers aiming to gain an understanding of test-takers' actual processing (and thus on what processes underlie correct item responses) have so far often relied on concurrent or retrospective verbal reports produced by test-takers (e.g. Anderson, Bachman, Perkins & Cohen, 1991; Goa & Gu, 2008; Phakiti, 2003; Rupp, Ferne & Choi, 2006; Yamashita, 2003; Yi'an, 1998). Although the use of these methods has resulted in valuable data on test-takers' processing, these methods have also been criticized for reactivity and veridicality risks, i.e. changing the thought process or length because of the act of verbalising, and inaccurately reflecting the process due to omissions or additions (Barkaoui, 2011; Ericsson & Simon, 1993).

Recent initiatives have started exploring the use of eye-tracking technology in the field of language testing as a tool for item validation, with promising initial findings (e.g., Bax, 2013; Bax & Weir, 2012; Gorin, 2006; McCray, 2013; McCray, Alderson & Brunfaut, 2012). However, to the researchers' knowledge, no guidelines exist on how to apply the technology to best fulfil this function. In methodological experimentation with eye-tracking technology to look into reading test items, McCray, Alderson and Brunfaut (2012) found that a combination of retrospective interviews, with eye-tracking traces providing the stimulus, proved to generate rich data on test-takers' cognitive processes, a position affirmed by Holmqvist et al. (2011). This study, therefore, aimed to further explore the utilisation of eye-tracking, and also the synergy of retrospective interviews with eye-tracking traces providing the stimulus, to look into cognitive processing. It was thought that the two data sources would allow the triangulation of information on test-takers' cognitive processes during reading item completion. At the same time, empirical evidence of these cognitive processes could potentially provide insights into the relationship between what aspect of the construct items intended to measure and what empirical data imply they are measuring, which constitutes crucial information for Aptis validation purposes.

When reading, our eyes make a series of small jumps along the lines, called *saccades*, rather than moving continuously across the line (Rayner, Juhasz & Pollatsek, 2007; Rayner, Pollatsek, Ashby & Clifton, 2012). At the end of each saccade, the eyes rest or fixate on a point on the page for a very brief instance; these pauses are termed *fixations*. Our eyes can jump over one or more words at a time, without fixating on every word of a text. In fact, it has been shown in L1 reading research that shorter or predictable words are often skipped or 'jumped over' (Blanchard, Pollatsek & Rayner, 1989; Brysbaert & Vitu, 1998; Ehrlich & Rayner, 1981; Rayner & Well, 1996). The reason our eyes make these saccades is to bring text into regions of higher visual resolution (termed 'acuity') for processing. This relates to the fact that our visual field varies in degree of acuity, i.e., it is more 'blurred' the further away from our fixation centre. Importantly, Rayner (1998) suggested that there is a close link between the point in a text at which our eyes have the highest acuity, the focal region, and the object of our attention at that time.

The length of the saccades our eyes make varies. For example, research has found that, as the cognitive load increases, the saccade length decreases when doing tasks such as driving or counting (Ceder, 1977; May, Kennedy, Williams, Dunlap & Brannan, 1990; Recarte & Nunes, 2003; Troy, Chen & Stern, 1972). Such effects have been argued for the context of reading too. For instance, in English-L1 reading, saccade length has been found to differ depending on the type of reading, with mean saccade sizes of 2° for silent reading versus 1.5° for reading aloud (Rayner, 1998, p. 373), reflecting that we cannot speak as fast as we can read (Levy-Schoen, 1981).

Furthermore, Rayner and Pollatsek (1989) stated that the saccade lengths of children who are learning to read are the same as the distance between each letter on the page, and that this reflects the fact that they are processing each letter during word recognition. They also stated that the average saccade lengths of weaker readers are lower, presumably reflecting these readers' difficulties at lower levels of reading processing. As the efficiency of lower-level processing improves, so does general reading ability, and this is reflected in comparatively longer saccade lengths which allow better readers to do greater amounts of processing by taking in more information with each fixation of the eyes.

In English, most of the saccades will follow a left to right movement, i.e. the normal reading direction. However, around 10–15% of eye movements in reading are thought to be backward movements from right to left, termed *regressions* (Rayner et al., 2007, 2012; Rayner, 1998). A possible explanation for smaller regressive saccades is correcting for 'overshooting' the area of visual acuity most efficient for the processing of information. Larger regressive saccades of more than 10 character spaces (Rayner, 1998) might be a consequence of a breakdown in comprehension and an attempt to remedy the situation. Indeed, it has been documented that as text difficulty increases, the fixation length and number of regressions increase, while saccade length decreases (Blanchard et al., 1989; Jacobson & Dodwell, 1979). Also, backward movements made inside words are thought to be more representative of lexical activation processes such as Khalifa and Weir's (2009) word recognition and lexical access processing. Regressions between words, on the other hand, are understood to reflect mainly sentence integration processes such as syntactic parsing, establishing propositional meaning and inferencing as specified by Khalifa and Weir (2009) (Holmqvist et al., 2011). In addition, it has been found that as reading ability increases, the number of regressions made decreases (Holmqvist et al., 2011; Murray & Kennedy, 1988).

After a saccade, the amount of time a reader spends fixating on a particular area of a text is called the *fixation duration*. This is probably the most frequently used measure in eye-tracking research (Holmqvist et al., 2011). There are various subdivisions of this measure thought to be representative of different aspects of cognitive processing. L1 reading studies, for example, have shown that the amount of time a reader fixates on a word is shorter if the word is easier to locate in their mental lexicon and is coherent with the rest of the text (Clifton, Staub & Rayner, 2007).

Several more elements have been found to affect fixation duration, including corpus-derived word frequency, word familiarity, lexical ambiguity, morphological effects, contextual constraints and plausibility. For instance, Inhoff and Rayner (1986) found that, even when controlling for the effect of longer words tending to be less frequent, infrequent words had higher fixation times. Williams and Morris (2004) found that words that were more familiar, as judged by participants, were associated with lower fixation duration. Words that are semantically ambiguous (e.g. (financial) bank and (river) bank) or phonologically ambiguous heteronyms (e.g. polish, minute or wind) have been shown to lead to longer fixation times, dependent on disambiguating contextual information contained in the text (Serenó, O'Donnell & Rayner, 2006). Similarly, if a word is less predictable from the information provided by the preceding words, the fixation duration tends to be longer (Ashby, Clifton & Rayner, 2005; Rayner, Ashby, Pollatsek & Reichle, 2004), and if an implausible word is located in a sentence, the fixation duration on that word is greater (Rayner, Warren, Juhasz & Liversedge, 2004). In addition, within-word morphological effects have been shown to have an influence on fixation times in studies by Andrews, Miller and Rayner (2004) on English, and by Hyönä and Pollatsek (1998) and Pollatsek, Hyönä and Bertram (2000) on Finnish. Also, as with saccade lengths, mean fixation times have been found to differ according to the type of reading; silent L1 reading was found to have a mean fixation time of 225ms, whereas reading aloud has a mean fixation time of 275ms (Rayner, 1998, p. 373). These differing fixation lengths reflect the nature of the task. The three key eye-tracking measures described above – saccades, fixations and regressions – are visually illustrated in Figure 2.

Figure 2: Key eye-tracking measures



Apart from gaining insights by looking into saccades, regressions and fixations, eye-tracking researchers have also analysed readers' eye traces according to what are called *areas of interest* (AOIs) – specific parts of the stimulus presented on the screen that the researcher is interested in (Holmqvist et al, 2011). This could, for example, be a text's title or individual paragraphs, or, in the context of testing reading, the text versus the items. Analyses, thereby, typically focus on measures of the time readers spend within an AOI and the transitions readers make between AOIs. In a study investigating cognitive processing during the completion of banked gap-fill items, McCray (2013), for example, found that better performing test-takers made fewer visits to the 'bank of words' needed to complete the text than did lower performing test-takers.

So far, the majority of studies using eye-tracking technology to look into reading were carried out on L1 readers. Fewer insights have been gained in this manner into L2 reading, and most who used the technology focussed on L2 sentence parsing and ambiguity resolution (Dussias & Sagrara, 2007; Dussias, 2003, 2010; Frenck-Mestre, 2002; Keating, 2009; Roberts, Gullberg & Indefrey, 2008). Similarly, little research in language testing has made use of eye-tracking methodologies, despite the insightful findings in L1 research. However, if we accept Rayner's (1998) position that there is a link between the position of a reader's gaze and the object of the reader's attention, then the detailed temporal information that can be gleaned from eye-movement recordings is potentially useful for unravelling test-takers' processing during task completion and researching item validity. Even at a more basic level, eye-tracking data could prove valuable. For example, in one of the few studies conducted in language testing, Gorin (2006) was able to detect construct-irrelevant variance by inspecting a scan path (a visual representation of a test-taker's eye movements across a text); one of the test-takers had been able to answer the reading item without looking at the text. Although potentially this type of conclusion can also be drawn from verbal reports, the eye-tracking data provided incontestable, direct evidence in this case. In fact, the uninterrupted reading while eye-tracking is more naturalistic than other techniques whereby test-takers need to combine their language processing with simultaneous verbalisations of that processing.

Nevertheless, as mentioned earlier, eye-tracking studies in language testing (Bax, 2013; Bax & Weir, 2012; McCray, 2013; McCray, Alderson & Brunfaut, 2012) have found the combined use of eye-tracking with a form of retrospection particularly helpful. In these studies, test-takers' eye traces were presented to them after item or task completion, as a stimulus to remind them of their processing. The eye-trace videos thus enabled the collection of stimulated recalls (Gass & Mackey, 2000). At the same time, the studies concluded that the verbal report data gathered in this manner triangulated the interpretations made on the basis of the eye-tracking data analyses, and gave the researchers more confidence in their conclusions.

3. RESEARCH QUESTIONS

The general aim of this study, as stated in Section 1, was translated into the following overarching research question:

RQ1. What cognitive processes do test-takers employ during Aptis reading task completion?

To gain more detailed insights into test-takers' cognitive processing, three sub-questions were formulated, exploring the nature of cognitive processing depending on task and test-taker characteristics. The first sub-question (RQ1a) aimed at examining cognitive processing while completing the tasks associated with each of the four target CEFR levels (A1 to B2).¹ In terms of test-taker variables, differences were examined in cognitive processing depending on test-takers' L2 reading proficiency (RQ1b) and their overall L2 proficiency (RQ1c).

RQ1a. Are there any differences in cognitive processes between items targeting different CEFR levels (and the associated task types)?

- CEFR A1 target, multiple choice (MC) gap-fill
- CEFR A2 target, sentence ordering
- CEFR B1 target, banked gap-fill
- CEFR B2 target, matching headings

RQ1b. Are there any differences in cognitive processes depending on test-takers' L2 *reading* proficiency, as measured by the Aptis reading component?

RQ1c. Are there any differences in cognitive processes depending on test-takers' *overall* L2 proficiency, as measured by all different Aptis test components?

4. METHODOLOGY

Based on a literature review (see Section 2) and previous research experience (McCray, 2013; McCray, Alderson & Brunfaut, 2012), it was decided to combine eye-tracking and stimulated recall methodology to obtain data on test-takers' cognitive processes during Aptis reading task completion.

4.1 Participants

The participants in our study were 25 English as a Second Language (ESL) speakers from three different first language backgrounds: 10 were Thai-L1, 10 were Chinese-L1, and 5 were Russian-L1 speakers. With regards to gender, 44% were male and 56% were female. Their ages ranged between 18 and 40 years old ($M=23.9$). Of the participants, 20% were enrolled on a pre-session English language course, 32% were undergraduate and 48% were postgraduate students at a British university. They had been living in English-speaking countries for between six months and seven years ($M=1.6$ years).

¹ It should be noted that each target CEFR level in the Aptis test system is associated with a different task type. Any potential differences in cognitive processes may thus be due to task type, target CEFR level, or the combination of these two. Nevertheless, the cognitive processing at each individual CEFR level (and associated task type) can be looked into and described (but care has to be taken in explaining these).

4.2 Materials

4.2.1 Reading tasks

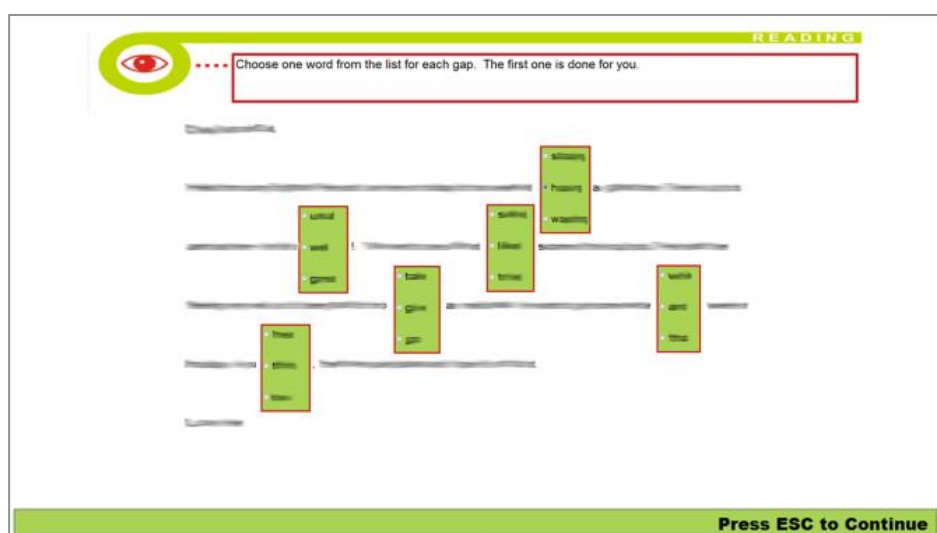
The test examined was the reading package of the computer-based Aptis test system developed by the British Council. The reading component constitutes a 30-minute test, consisting of four tasks which are of different task types and which each target a particular CEFR level and different aspects of reading comprehension (British Council, 2013). An overview is given in Table 1. Sample tasks are accessible on: <http://www.britishcouncil.org/aptis-practice-tests/AptisReadingPractice/>

Table 1: Aptis reading component structure


Part	Reading	Task type	CEFR level
1	Sentence comprehension	Multiple-choice gap-fill	A1
2	Text cohesion	Sentence re-ordering	A2
3	Short-text comprehension	Banked gap-fill	B1
4	Long-text comprehension	Matching headings	B2

Because of the specific research methodology of the study, the tasks were slightly re-formatted. Although the Aptis reading test is computer-delivered, the tasks could not be run online as that would not have allowed pausing and replays of eye traces for the purpose of the stimulated recall. Thus, the reading tasks needed to be transferred to a format compatible with the eye-tracker software. A second reason for re-developing the reading tasks' layout was that, to enable the interpretation of the eye traces, it needed to be clear at every point what the participant was looking at. Thus, features such as drop-down menus that can be made fully visible or invisible by clicking, and that overlay the underlying text, interfered with interpreting the link between a particular eye trace and what the participant was looking at (i.e. the underlying text or an aspect of the text in the drop-down menu). By means of two pilot studies with four participants (reported on in more detail in Brunfaut & McCray, 2014), the layouts were designed as shown in Figure 3.

Figure 3: Adapted Aptis reading task layouts



LOOKING INTO TEST-TAKERS' COGNITIVE PROCESSES WHILE COMPLETING READING TASKS
BRUNFAUT AND McCRAY



READING

Order the sentences below to make a story. The first one (1) is done for you.

1	The first sentence is done for you.
2	The second sentence is done for you.
3	The third sentence is done for you.
4	The fourth sentence is done for you.
5	The fifth sentence is done for you.
6	The sixth sentence is done for you.
7	The seventh sentence is done for you.
8	The eighth sentence is done for you.

READING

Read the text and complete each gap with a word from the list at the bottom of the page.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....


.....

.....

.....

.....

Press ESC to Continue



READING

Read the passage below quickly. Choose the most appropriate heading from the selection. There is one more heading than you need.

Mission to Mars

1. The first mission to Mars is planned for 2025. It will be a one-way mission, meaning that the crew will not return to Earth. The mission will be led by NASA and will consist of a crew of four astronauts. The mission will be the first time that humans have traveled to another planet.

2. The mission will be a one-way mission, meaning that the crew will not return to Earth. The mission will be led by NASA and will consist of a crew of four astronauts. The mission will be the first time that humans have traveled to another planet.

3. The mission will be a one-way mission, meaning that the crew will not return to Earth. The mission will be led by NASA and will consist of a crew of four astronauts. The mission will be the first time that humans have traveled to another planet.

4. The mission will be a one-way mission, meaning that the crew will not return to Earth. The mission will be led by NASA and will consist of a crew of four astronauts. The mission will be the first time that humans have traveled to another planet.

5. The mission will be a one-way mission, meaning that the crew will not return to Earth. The mission will be led by NASA and will consist of a crew of four astronauts. The mission will be the first time that humans have traveled to another planet.

6. The mission will be a one-way mission, meaning that the crew will not return to Earth. The mission will be led by NASA and will consist of a crew of four astronauts. The mission will be the first time that humans have traveled to another planet.

7. The mission will be a one-way mission, meaning that the crew will not return to Earth. The mission will be led by NASA and will consist of a crew of four astronauts. The mission will be the first time that humans have traveled to another planet.

Answering questions
1. _____
Answered correctly
2. _____

A different view of the world
3. _____
Answered correctly
4. _____

Only an experiment
5. _____
Answered correctly
6. _____

Installing the system
7. _____
Answered correctly
8. _____

Press ESC to Continue

As compared to the task layouts of the official Aptis reading test (see British Council, 2013, pp. 12–16), the following accommodations were made for the study's purposes.

1. Task type 1, *multiple-choice gap-fill*: The original task layout makes use of drop-down multiple-choice menus. To keep the task format as similar as possible to the original, the space between lines was increased so there was no overlap between the MC-options and lines of text beneath or above the line of the MC-gap. In addition, the MC-options' visibility was made permanent (rather than a clickable drop-down) to be able to link eye traces to a particular option.
2. Task type 2, *sentence re-ordering*: The original task layout requires the test-takers to move around sentences to drop them into the right sentence position to form a continuous text. However, as sentences can change positions, this does not allow linking eye-tracking data to a fixed piece of underlying text and thus makes interpretation of eye traces extremely difficult. Therefore, the task layout was re-designed to consist of fixed sentence positions and a space to the left-hand side to fill in the number of the text line the sentence would take in a continuous piece of text.
3. Task type 3, *banked gap-fill*: The layout of this task was kept similar to the original layout.
4. Task type 4, *matching headings*: The original task layout allows the test-taker to scroll the text up and down, with only a small part of the text visible at any one point in time. Also, the headings have to be chosen from drop-down menus that overlay other drop-down menus. To be able to link the eye-tracking data to what the participant was looking at, the text and options need to be visible at all times. Therefore, the layout of this task type was adapted so that the entire text is shown on the screen. In addition, the heading options are presented in a bottom box, and the participant needs to enter the number of the matching paragraph, instead of choosing the matching heading from a drop-down menu.

Two versions of the Aptis reading component, provided by the British Council, were administered to each participant, totalling eight reading tasks and 50 items.

4.2.2 Full Aptis test

Since the study also aimed to explore potential differences in cognitive processes depending on test-takers' L2 reading proficiency and overall L2 proficiency (RQs1b & 1c), participants' overall and L2 reading proficiency needed to be measured. In practice, the full computer-based Aptis test system – consisting of the components grammar/vocab, reading, listening, writing, and speaking – was used for this purpose.^{2,3} The test and the administration procedures, as stipulated by the British Council, were technically and operationally piloted with two participants.

² Note that the reading task versions used in the eye-tracking/stimulated recall study differed from those in the full Aptis test.

³ The role of the grammar and vocabulary component in the full Aptis system is described as follows by the British Council: "Aptis takes a unique view on how to deal with test error (standard error – a feature of all tests). Candidates near a decision point (e.g. the border between B1 and B2) will have their score on the grammar and vocabulary paper taken into account when a final CEFR grade is awarded". (<http://www.britishcouncil.org/aptis/packages>; retrieved on 15 September 2014)

4.3 Data collection methodology and procedures

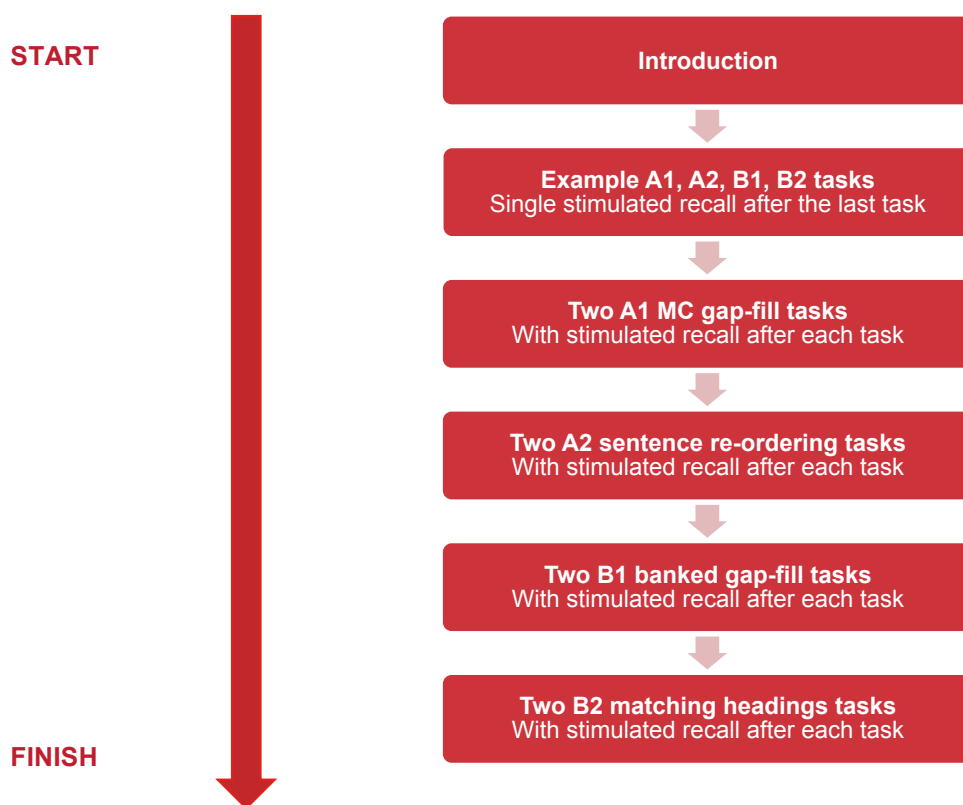
Phase 1

The data were collected from the participants over two sessions. During the first session, the participants completed the Aptis reading tasks while their eye traces were being recorded. This was immediately followed by a retrospective interview on participants' cognitive processes during task completion in the participant's L1, using their eye traces as stimuli for retrospection (also called 'stimulated recall methodology'). The aim of this first phase was to collect data that would inform the answer to the overarching research question (RQ1 *What are test-takers' cognitive processes during Aptis reading test completion?*) and the first sub-question (RQ1a *Are there any differences in cognitive processing while completing each of the four CEFR-linked task types?*).

Through a piloting process conducted with four pilot study participants, the procedure shown in Figure 4 was developed for this first part of the data collection. Factors that influenced this design included the training needs of the participants, the total time needed to complete the experiment, the concentration demands for the participants, and technical practicalities of the eye-tracking software (for more details see Brunfaut & McCray, 2014). At all times, decisions were made in light of the aim to gather high-quality data on test-takers' cognitive processes.

Given the choice of the data collection methods, the data in this first phase were collected from one participant at a time. The time of the sessions ranged between approximately 1 hour 30 minutes and 2 hours.

Figure 4: Flowchart of the first data collection session



As can be seen in Figure 4, the session started with an introduction, which included an explanation reminding the participant of the nature of the study, signing an ethical consent form, completing a participant background questionnaire, and a technical eye-tracking suitability test. If the suitability test was successful,⁴ the eye-tracking and stimulated recall procedure was started.

For each of the four Aptis reading task types, first an example task (provided to the researchers by the British Council) was shown to the participant. This allowed the participant to become familiar with, and try out, each task type. During the last sample task – matching headings – the stimulated recall procedure was trialled with the participant. Namely, the participant's eye traces were recorded during the sample task completion and then replayed to the participant, who was asked to recall and verbalise his/her task completion process. If necessary, feedback was given to the participant on the nature or quality of the stimulated recall example. The restriction to one stimulated recall trial (one task type, seven items) was based on experiences and feedback from the pilot study participants and the researchers' observations.

After the example tasks, the main data were collected. The participant was asked to complete two tasks of each CEFR-linked item type, which were presented according to increasing target CEFR-level (similar to the official Aptis reading component's structure). The participant's eye traces were simultaneously recorded. Each completed task was followed by a replay of the eye traces (pausing after each item completion attempt) and a request to verbalise how they had approached the reading task and items in general, what they had been thinking during task completion, and how they had arrived at each of their answers. For consistency, all stimulated recalls were conducted following a script with instructions and questions asked by the researcher. The script and procedures had been tried out and positively evaluated during the pilot study.

It should be noted that the stimulated recalls were audio- and video-recorded. During piloting, it was observed that participants also pointed at the screen when recollecting their cognitive processes by means of the eye traces replay and task stimuli. To facilitate the understanding and interpretation of the stimulated recall data at the analysis stage, it was decided to also visually capture the stimulated recall process.

To enable the participants to express their thoughts with ease, the stimulated recalls were conducted in the participant's first language (L1), with the option of using English if the participant wished to do so. In order to allow for L1 stimulated recalls, at the research design stage we had identified three groups of L1 backgrounds for which we would be able to recruit a sufficient number of participants. Also, we chose more than one group to represent some L1 variation and better reflect variation in the official Aptis test-taking population. An additional consideration was the availability of suitable research assistants who would be able to conduct the stimulated recalls in the participants' L1, and transcribe and translate the recalls.

In practice, we approached Thai-L1, Chinese-L1 and Russian-L1 ESL speakers for participation in our study, and we recruited research assistants from the same three first language backgrounds. The research assistants were all linguists, specialised in language testing and second language acquisition (SLA), with experience in collecting verbal protocols. They were given a two-hour training session in the technical use of the eye-tracking software and audio- and video-recording hardware, as well as in the practical use of the stimulated recall methodology and overall procedures. These research assistants had also been involved in the pilot study as participants in order to familiarise themselves with the materials and procedures, and to experience the study from a participant's point-of-view. All data collection sessions were supervised by one of the two main researchers.

⁴ Three volunteers had to be turned down due to problems recording their eye traces in a suitable manner for data analyses. This is typically due to issues such as long eyelashes, droopy eyelids, or varifocal glasses (Holmqvist et al., 2011).

Phase 2

During the second data collection session, the same participants who had taken part in the eye-tracking/stimulated recall session were administered the full Aptis test (all five components). This was done to obtain a measure of the participants' English reading proficiency and of their overall English language proficiency. The combination of the eye-tracking/stimulated recall data (from Phase 1) with the Aptis results (Phase 2) was necessary to be able to analyse potential differences in cognitive processes depending on test-takers' L2 reading proficiency (RQ1b) and depending on their overall L2 proficiency (RQ1c).

The full Aptis test was administered in small groups (depending on the participants' availability) in a computer lab at the researchers' institution. The session strictly adhered to the official Aptis test's procedures and was supervised by one of the researchers. Participants took approximately 1.5 to 2.5 hours to complete the full Aptis test (the maximum set by the Aptis system is three hours), with an optional break between components.

4.4 Ethical procedures and consent

In line with the regulations at the researchers' institution, ethical approval was sought and obtained. The participants were provided with a written information sheet detailing the nature of the study, their involvement, and the contact details of the researchers and their Head of Department. The researchers also orally explained to the participants the study and the required involvement. All participants gave their consent in writing.

4.5 Data analyses

In the context of this study, with its relatively complex stimuli in terms of the target range of cognitive processes (British Council, 2013), the eye-tracking and stimulated recall methodologies are mutually complementary in the analysis of the test-takers' cognitive processing in reading. It was anticipated that the eye-tracking analyses would be particularly useful to explore lower-level processing, while the stimulated recall analyses would generate more insights into higher-level processing. For example, stimulated recalls can tell us little about a lower-level processing variable such as word recognition speed, while eye-tracking analyses are unlikely to provide information on inferences made by the test-taker. The analyses conducted on both data sources – eye movements and stimulated recalls – are described in the following sections.

4.5.1 Eye-tracking analyses

Measures

A total of 11 eye-tracking metrics were looked into, derived from the test-takers' fixations, saccades and regressions, and motivated by their use in past research (see Holmqvist et al., 2011) and their potential process indication characteristics (see hypotheses below). Because this study concerns the testing of reading (rather than reading *per se*), test-takers' processing and eye traces on the text were distinguished from those on the items. As such, the eye-tracking measures could be subdivided into three processing-type groups: a) global processing metrics, b) text processing metrics, and c) task processing metrics.

The global processing measures constitute more summative measures of task completion and are taken from data pertaining to both the text and response options of an item. Text processing measures are specific to the text of the item and are thus taken from data pertaining to the written text of the item and not the response options. Task processing measures are those which pertain to the interactions between the text and the response options, or contrast measures on the text with measures on the response. An overview of the measures, as well as their technical definitions, is provided in Table 2.

Table 2: Eye-tracking metrics

Processing focus	Measure	Technical definition
Global processing	Total number of fixations	The sum of the number of fixations as defined by the fixation filter.
	Total fixation time on text and responses	The sum of all fixation durations on text and response, expressed in seconds.
Text processing	Number of forward saccades*	A forward saccade is a movement between two fixations, as defined by the fixation filter, from point x to point y where point y lies to the right of point x and is within plus or minus 10 degrees horizontally.
	Median length of forward saccades*	Median length, expressed in pixels, of all forward saccadic movements.
	Number of regressions	A regression is a movement between two fixations, as defined by the fixation filter, from point x to point y where point y lies to the left of point x, is within plus or minus 10 degrees horizontally, and is below some defined threshold designed to stop line returns being classified as regressions.
	Median length of regressions*	Median length, expressed in pixels, of all regression movements.
	Proportion of regressive movements	The number of regressive movements divided by the sum of all eye movements (i.e. the number of forward saccades and the number of regressions).
	Median fixation duration*	The median of the fixation durations, expressed in milliseconds.
	Sum fixation time on text per word	The sum of the fixation time on the text, measured in seconds, divided by the number of words in the text of the item.
Task processing	Proportion of time spent fixating on response options	The total fixation time on response options divided by the total fixation time on the text and response options.
	Number of Aol switches between text and response options	The number of movements between Areas of Interest (Aols) containing text and an Aol containing the response options.

*Scaled for font size (see below).

Fixation filter

A number of methods exist to determine what constitutes a fixation. Therefore, it has been considered important in eye-tracking research to report the fixation filter – the algorithm for detecting fixations – that has been used, as different filters may lead to different results (Holmqvist et al., 2011). In this study, a velocity and acceleration-based filter was chosen since this functions well on high-speed eye-trackers such as the Tobii 300, which was used in this case. This type of filter uses the speed and acceleration of the eye during a saccade to determine a fixation, and thus requires high temporal resolution for accurate fixation detection since saccades only last a fraction of a second. More specifically, the Tobii I-VT filter with its default settings was adopted based on an assessment of the effect of different algorithms and settings on the quality of the data.

Font scaling

In order to fit the texts and items of each task on a single slide for presentation on the eye-tracker, the font sizes had to be varied across task types. To allow for comparisons across CEFR-linked task types, some measures were scaled. Namely, distance measures, which would reduce with a smaller font size (indicated with * in Table 2), were scaled by 1.07 for the task type 'banked gap-fill' and by 1.27 for the task type 'matching headings'.

Analyses

In order to investigate test-takers' processing while completing each of the CEFR-linked task types (RQ1a), the eye-tracking data collected on the items of each CEFR-linked task type were analysed according to the 11 metrics defined in Table 2. Potential differences depending on target CEFR level (with the confounding factor of task type) were explored by means of Kruskal-Wallis tests. This was followed by subsequent pairwise comparisons to compare measures between CEFR-level pairs.

Non-parametric tests were chosen for the analysis, since suitable normal transformations could not be found for many of the measures. It should be noted that no suitable effect size statistic exists for the Kruskal-Wallis test (Field, Miles & Field, 2012). No corrections were made to the p-values based on the multiple tests on the same dataset as these are known to reduce statistical power via the inflation of type one error rates (Field et al., 2012), and the statistical power is already limited on the relatively small sample.⁵

For each of the eye-tracking measures, a hypothesis was formulated on the direction of the measure and processing in relation to CEFR level (RQ1a). These are presented in Table 3.

To explore differences in test-takers' cognitive processing depending on test-takers' English L2 reading proficiency (RQ1b) and their overall English L2 proficiency (RQ1c), Spearman correlations were run. These analyses were conducted to compare students' overall reading test completion processing in relation to their L2 (reading) proficiency, and also their performances per CEFR-linked task type in relation to their L2 (reading) proficiency.

For each of the eye-tracking measures, a hypothesis was formulated on processing in relation to test-takers' (reading) proficiency (RQ1b & RQ1c). These are presented in Table 4.

⁵ It should be noted that collecting good quality eye-tracking data is complex and costly, hence the difficulty to reach large sample sizes.

Table 3: Eye-tracking measure hypotheses in relation to reading task CEFR level (RQ1a)

Processing focus	Measure	Hypothesis with reference to RQ1a
Global processing	Total number of fixations	As the CEFR level of the tasks increases, the total number of fixations will increase. This is directly due to the fact that the texts for the harder tasks are longer. Longer texts test higher-level cognitive processes which relate to comprehension at the sentence level and above.
	Total fixation time on text and responses	As the CEFR level of the tasks increases, the total fixation time on the text and responses will increase. This measure is closely linked with 'total number of fixations', but it is more sensitive to the total time spent on the text and less sensitive to movement around the text.
Text processing	Number of forward saccades	As the CEFR level of the tasks increases, the number of forward saccades will increase. This measure represents the fact that higher-level tasks have longer and more complex texts that require more processing.
	Median length of forward saccades	As the CEFR level of the tasks increases, the median length of a forward saccade will decrease. This is due to the increased cognitive load on the test-taker as a function of CEFR level (higher level, higher cognitive processing load).
	Number of regressions	As the CEFR level of the tasks increases, the total number of regressions will increase. This is due to two factors: firstly, a regression is more likely in a longer text; and secondly, regressions are more likely as the text becomes more challenging.
	Median length of regressions	As the CEFR level of the tasks increases, the median length of the regressions will increase. This relates to the notion that more complex texts in the higher-level CEFR tasks will generate more between-word regressions, as they are designed to measure higher-level cognitive processing, than the lower CEFR level tasks.
	Proportion of regressive movements	As the CEFR level of the tasks increases, the proportion of regressive movements will decrease. As the texts increase in complexity, there will be a greater need to perform regressions in order to facilitate comprehension.
	Median fixation duration	As the CEFR level of the tasks increases, the median fixation duration will increase. This would be due to the test-takers requiring longer fixations to comprehend the more complex texts and perform the more complex operations required by the higher-level tasks.
	Sum fixation time on text per word	As the CEFR level of the tasks increases, so does the proportion of time spent on the text per word. This would be due to the fact that the increasing cognitive demands placed on the test-takers by the higher level texts require greater processing per word.
Task processing	Number of Aol switches between text and response options	As the CEFR level of the tasks increases, the number of switches between the text and the responses will increase. This would be due to the increasing difficulty in integrating the information contained in the text with the response in the selection of the correct answer.
	Proportion of time spent fixating on response options	As the CEFR level of the tasks increases, the proportion of time spent fixating on the responses will decrease. This would be due to the proportionally increasing demand of the text over the responses as the level of the tasks increases.

Table 4: Eye-tracking measure hypotheses in relation to L2 reading proficiency (RQ1b) and overall L2 proficiency (RQ1c)

Processing focus	Measure	Hypothesis with reference to RQ1b and RQ1c
Global processing	Total number of fixations	As the ability of the test-taker increases, the number of fixations required to complete a task will decrease. This reflects the increased processing efficiency of higher ability test-takers who are able to process the text with fewer fixations, i.e. they have fewer breakdowns in comprehension leading to re-reading text, they use longer saccades (thus fewer fixations) to process text and/or they find the correct response quickly, and are confident in their selection, without the need for extensive searches or validation of their response.
	Total fixation time on text and responses	As the ability of the test-taker increases, the amount of time it takes fixating on a task will decrease. This reflects the increased processing efficiency of higher ability test-takers (see 'Total number of fixations').
Text processing	Number of forward saccades	As the ability of the test-taker increases, the number of forward saccades on the text of a task will decrease. This reflects the increased processing efficiency of higher ability test-takers (see 'Total number of fixations').
	Median length of forward saccades	As the ability of the test-taker increases, the median length of forward saccades on the text will increase. This is due to more skilful readers being able to process more information during each fixation.
	Number of regressions	As the ability of the test-takers increases, the number of regressions will decrease. This is because higher-ability test-takers need to solve fewer processing issues.
	Median length of regressions	As the ability of the test-takers increases, so will the length of regressions. This is because higher ability test-takers have fewer problems with word recognition and lexical access and, thus, perform fewer shorter regressions.
	Proportion of regressive movements	As the ability of the test-taker increases, the proportion of regressive movements will decrease. This would be due to the effect of poorer test-takers' need to re-read sections of the text to facilitate comprehension.
	Median fixation duration	As test-taker ability increases, the median fixation duration will decrease. This would be due to the better readers processing the information at each fixation faster than the poorer readers.
	Sum fixation time on text per word	As the ability of the test-takers increases, the sum fixation time per word will decrease. This reflects the ability of the higher-level test-takers to process the information contained in the text more quickly than the lower level test-takers.
Task processing	Number of Aoi switches between text and response options	As the ability of the test-taker increases, the number of switches between the text and the responses will decrease. This would reflect the better test-takers being more able to process the text of the task and hold the representation in memory and not require extensive switching to the responses.
	Proportion of time spent fixating on response options	As the ability of the test-takers increases, the proportion of time spent fixating on the responses will increase. This would be due to the comprehension of the texts of the tasks presenting a proportionally smaller challenge to the better readers than to the poorer readers.

4.5.2 Stimulated recall analyses

The verbal reports produced by the participants were transcribed and translated from the participants' L1 into English by the research assistants (who were linguists specialised in English language testing and SLA). The transcriptions were done on the basis of the video recordings, while the audio-recordings served as a back-up in case the video sound was unclear or there were technical glitches in the recordings. The video data had the advantage of being able to link what participants said to what they saw or pointed at on the computer screen (which was particularly useful in cases where they used referents) and to add comments to the transcripts on visual aspects of the stimulated recalls and reading tasks.

The translated transcripts were uploaded in the qualitative data analysis software Atlas.ti v7, and coded by one of the main researchers. Khalifa and Weir's (2009) model of cognitive processing in reading, as represented in Figure 1, served as the basis of the coding framework. The main reasons for this choice were that this model was developed with the context of testing reading in mind and within Weir's (2005) approach to test validation, and that it has informed the design of the Aptis test. More specifically, given the present study's focus on cognitive processing, each of the processes specified in Khalifa and Weir's (2009) central core was adopted in the coding framework. In addition, the goal setting processes were also used as codes, since the Aptis reading test specifications stipulate particular types of reading for each of the CEFR-linked task types (see e.g. Dunlea, 2014). The resulting codes are listed in the first two columns of Table 5.

Four extra codes were arrived at during the coding process, on the basis of the nature of the data. Since it concerns reading in the context of a test with selected-response tasks, in a couple of instances, participants indicated that they decided on their answer purely by guessing rather than applying reading strategies, hence the category 'pure guess'. The category 'collocation' was added because in some cases participants did not seem to consider the options at hand or aim for comprehension, but said they completed the item based on collocational knowledge. In a limited number of cases, participants primarily relied on factual background knowledge to complete an item, hence the category 'background knowledge'. Finally, the category 'creating paragraph level representation' was added, specifically with reference to the fourth task type 'matching headings'. It was felt that this would allow a more precise description of the type of processing during completion of this task type consisting of a lengthy text with eight paragraphs. This additional category allows distinguishing between instances where test-takers create an overall text level representation versus a representation of an individual paragraph.⁶

Table 5: Coding framework stimulated recalls

Goal setting codes	Central core codes	Additional codes
Local reading	Word recognition	
Global reading	Lexical access	
Careful reading	Syntactic parsing	Collocation
Expeditious reading – scanning	Establishing propositional meaning	
Expeditious reading – skimming	Inferencing	Background knowledge
	Building a mental model	
	Creating text level representation	Creating paragraph level representation
	Creating intertextual representation	
		Pure guess

⁶ Note that in the other three task types, the text constitutes one paragraph only. The code 'text level representation' in those cases refers to that single paragraph, the entire text.

The same set of processes was used for coding items which the participants answered correctly and those they did not get right, but with the added qualification of 'W' for 'wrong' to the code name for the incorrectly answered items.

To determine what cognitive processes test-takers employ during Aptis reading task completion, the number of occurrences of each of the coding categories (Table 5) in the stimulated recalls of all test-takers was calculated. A distinction was thereby made between codings associated with correctly answered items and those associated with incorrectly answered items. This separate analysis was considered meaningful, because from a validation perspective, it is important to know whether or not the correctly answered items tested the intended aspects of reading.

To gain more insights into the data, several sub-analyses were conducted. Firstly, to explore differences in cognitive processing of items targeting different CEFR levels (RQ1a), the codings associated with tasks targeting a particular CEFR level were tallied separately. This allowed us to gain an understanding of cognitive processing for each individual set of CEFR-linked items, as well as to compare between CEFR-levels. Secondly, to investigate processing differences depending on test-takers' L2 reading proficiency (RQ1b), the stimulated recall data were split into groups according to the participants' L2 reading ability, expressed in CEFR proficiency levels. The latter had been measured by means of the reading component of the Aptis test (which did not contain any of the tasks used in the eye-tracked Aptis reading test versions). The stimulated recall codings were then analysed for each of the reading proficiency groups, and also compared across groups. Thirdly, the stimulated recall data were explored for cognitive processes associated with test-takers' overall L2 proficiency (RQ1c), whereby proficiency was established on the basis of test-takers' performance on the full Aptis test (the four skill components). The stimulated recall codings were then analysed for each L2 proficiency group separately, and comparisons between frequencies of code categories (average per test-taker, and corrected for the number of items) were made across groups.

All three types of sub-analyses mainly focussed on processes associated with correctly completed items, since these should reflect the intended processes to constitute a valid test. Nevertheless, code frequency analyses of the processes associated with incorrectly answered items were also conducted (but are not extensively reported on in this document), to inform the various sub-analysis findings. Similarly, cross-overs between CEFR level tasks and (reading) proficiency groups were done to add even more detailed insights into the sub-analysis results.

5. FINDINGS

5.1 Descriptive statistics

Table 6 describes participants' performance results on the Aptis readings tasks they completed while their eye movements were being recorded and on the basis of which they produced stimulated recalls. As can be seen from the mean scores and scoring range, overall, the participants performed well on the tasks.

Table 6: Descriptive statistics – Aptis reading tasks used for eye-tracking and stimulated recall (n=25)

	Max. score	Min	Max	M	SD
All tasks	50	32	48	40.56	4.53
A1 tasks	10	5	10	8.48	1.42
A2 tasks	12	5	12	9.88	2.42
B1 tasks	14	8	13	11.16	1.46
B2 tasks	14	6	14	11.04	2.21

Participants' results on the online, complete Aptis system were retrieved by the British Council from its database and provided to the researchers. For each participant, six different total scores were reported: 'all components' (total of the four skills), 'grammar and vocab', 'reading', 'listening', 'speaking', and 'writing'. Descriptive statistics are presented in Table 7. The fact that the performances on the reading component of the full Aptis (see third row in Table 7) were very similar to those on the Aptis reading tasks used for eye-tracking (see first row in Table 6) suggests that the eye-tracking methodology and task presentations did not considerably interfere with measuring participants' reading proficiency (as expressed in the score).

Table 7: Descriptive statistics – full Aptis system (n=25)

	Maximum score	Minimum	Maximum	Mean	Standard deviation
All skill components	200	142	184	166.72	12.89
Grammar and vocab	50	25	44	37.48	4.33
Reading	50	34	50	42.32	5.09
Listening	50	30	48	41.76	4.67
Speaking	50	25	48	39.56	5.41
Writing	50	35	48	43.08	4.03

The British Council also maps test-takers' performances on each of the four skill components to the CEFR. For each component, the number of participants assessed to be at a particular CEFR level are presented in Table 8. The table shows that, although we aimed to recruit L2 speakers in the range A2-C, the volunteers willing to participate were evaluated to be independent and proficient users (Council of Europe, 2001), particularly in terms of their receptive skills.

Table 8: Aptis components – CEFR levels of participants (n=25)

Aptis component	A1	A2	B1	B2	C
Reading	0	0	4	10	11
Listening	0	0	0	5	20
Speaking	0	1	5	17	2
Writing	0	0	4	12	9

5.2. Eye-tracking

To gain insights into test-takers' cognitive processing during Aptis reading test completion, participants' eye movements were analysed, as well as their thought processes reported by means of stimulated recalls. In this section, the results of the eye movement analyses are reported.

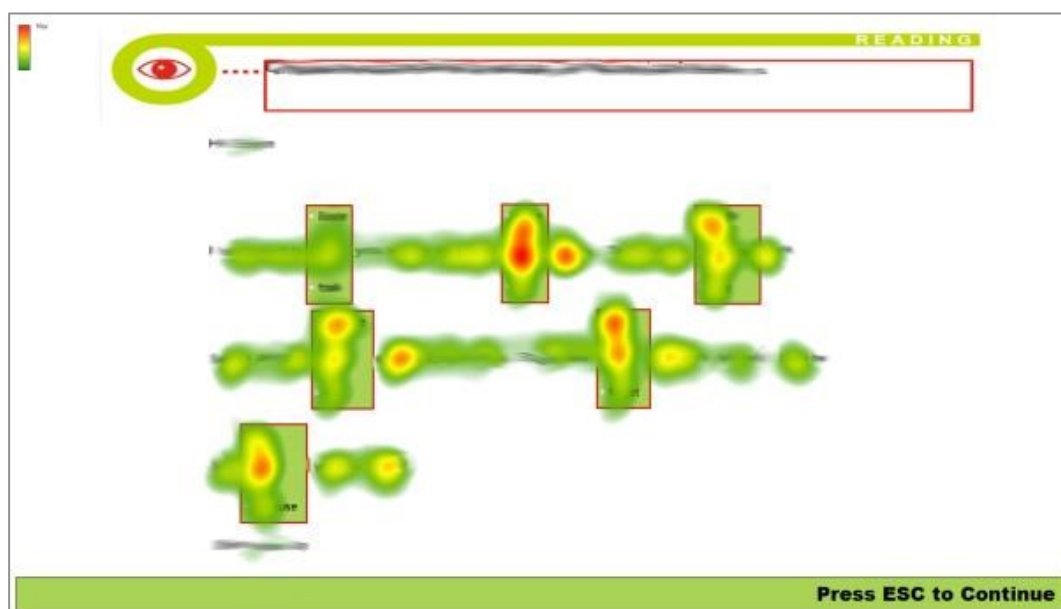
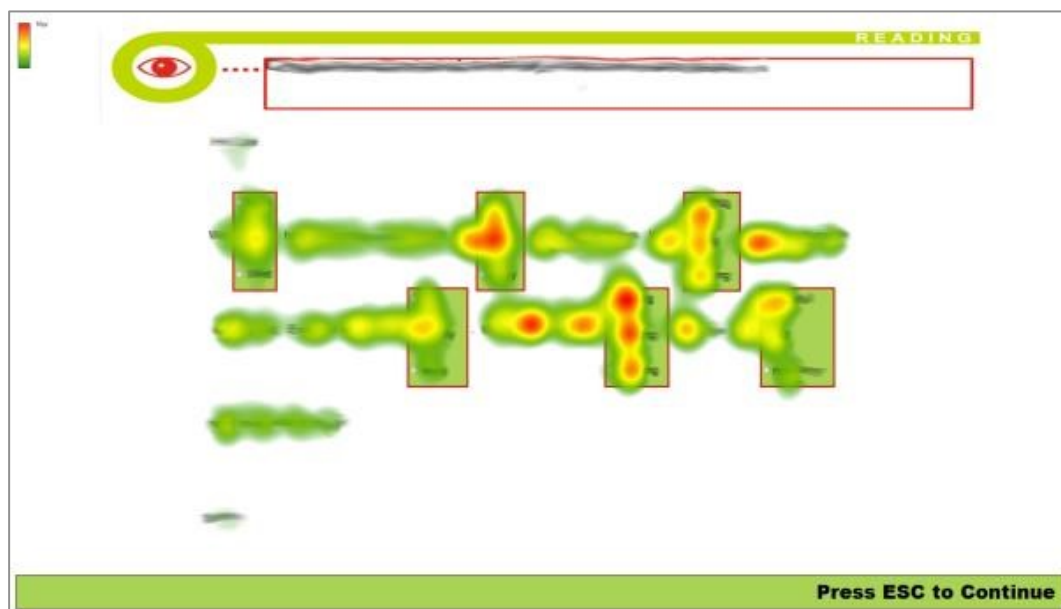
To obtain a first, overall understanding of test-takers' cognitive processing during Aptis reading test completion (RQ1), heat maps resulting from the recordings of participants' eye movement were inspected. Figure 5 shows these visualisations for all items per task. In heat maps, a range of colours from green through yellow to red are used to represent the aggregate amount of time a participant spends focusing on a particular area of the input. Areas of the input which receive no fixations remain transparent. To improve interpretability, the heat maps presented in Figure 5 have been plotted in such a manner that each participant's data has the same weight.

Clear patterns emerge from these visualisations. As can be seen in Figure 5, in the A1 tasks (MC gap-fill) the majority of the participants' attention was directed towards the multiple choice response options and the words surrounding the response options. This seems to suggest a more local, careful reading approach with an emphasis on lower-level processing such as syntactic and semantic parsing of the words surrounding/suggested for the gaps to select the correct response. For the A2 sentence ordering tasks, it can be seen that most attention is given to the beginning of the sentence options. Frequently, in the A2 items, the beginning of the sentences carry temporal adjectives, anaphors and logical connectors to be resolved, giving the test-taker key information to respond correctly. This may indicate a tendency to try to gain a more global picture of the texts and how different sentences fit together or follow one another. The heat maps of the B1 tasks, banked gap-fill tasks, show a similar picture to the A1 tasks, i.e. the attention of the test-takers was more directed towards the words surrounding the gaps. This again is likely to indicate elevated levels of more careful local considerations and lower-level processing while trying to complete the items.

In contrast, the B2 items show a different visual pattern. For the longer texts used in this 'matching headings' task type, considerable variance can be seen in the extent to which different paragraphs were read by the test-takers. Some paragraphs seem to have been read in their entirety, while other paragraphs only received little attention. Presumably, the more difficult items in the tasks required more careful global reading. Overall, there seems to be a tendency with the B2 tasks for the first few lines in a paragraph to be read more than the last lines. Potentially, if the test-taker had identified what they believed to be the correct answer after reading the first few lines, there was little reason to read on.

Figure 5: Heat maps of Aptis reading tasks

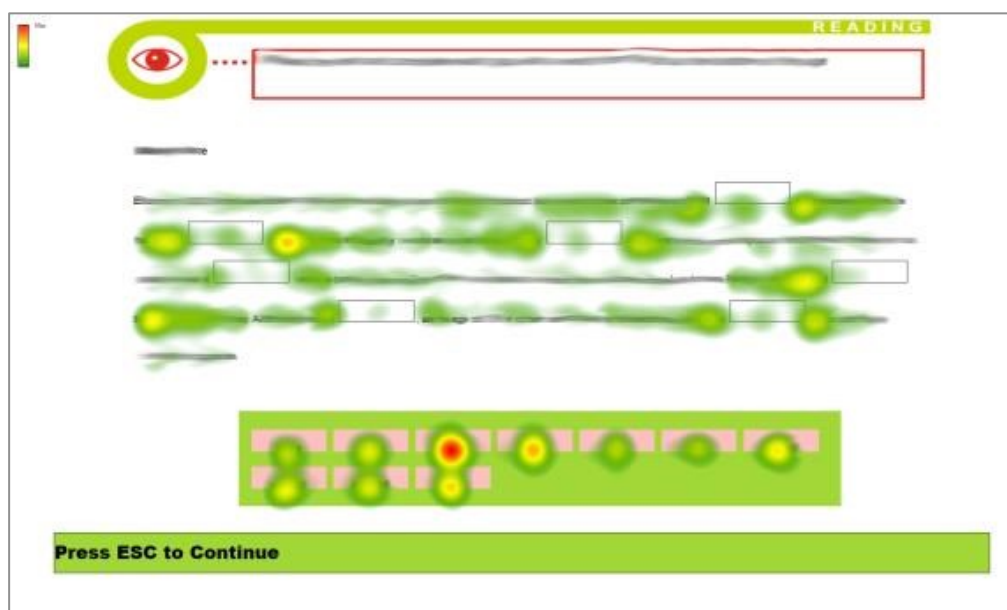
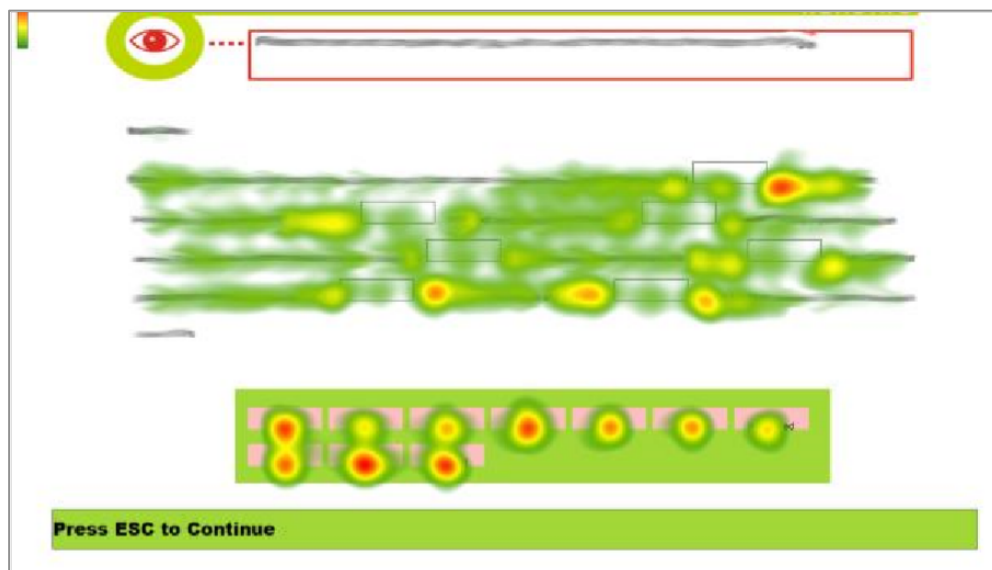
A1 tasks



A2 tasks



B1 tasks



B2 tasks



To explore these first, general impressions from the eye movement data in more depth, several sub-analyses were conducted, as described below.

5.2.1 Eye-tracking findings on cognitive processes when completing items targeting different CEFR levels (RQ1a)

The first sub-question (RQ1a) of the study was:

Are there any differences in cognitive processes between items targeting different CEFR levels (and the associated task types): CEFR A1 target/ MC gap-fill; CEFR A2 target/sentence ordering; CEFR B1 target/banked gap-fill; and CEFR B2 target/matching headings?

To inform the answer to this research question, the test-takers' eye movements were analysed per CEFR level of the tasks, in terms of the measures presented in Table 2. Descriptive statistics of the 11 measures are provided in Table 9 (IQR=Interquartile range).

Table 9: Descriptive statistics – Eye-tracking measures

		Global processing measures		Text processing measures							Item processing measures	
		Total number of fixations	Total fixation time on text and response (seconds)	Number of forward saccades	Median length of forward saccades (px)	Number of regressions	Median length of regressions (px)	Proportion of regressive movements	Median fixation duration (ms)	Sum of fixation time on text per word (seconds)	Number of Aol switches between text and responses	Proportion of time spent fixating on responses
All items	Median	3446	810	1806	68	855	-63	0.28	222	0.60	500	0.30
	IQR	882	248	364	14	491	16	0.07	18	0.16	177	0.03
	Min	2087	480	1143	53	393	-93	0.18	163	0.39	323	0.27
	Max	9429	2280	5011	100	2858	-45	0.37	261	2.00	1157	0.35
A1	Median	188	50	56	71	55	-69	0.36	220	0.64	71	0.45
	IQR	58	15	16	17	24	25	0.06	19	0.32	27	0.06
	Min	118	31	34	58	33	-110	0.22	161	0.41	41	0.35
	Max	447	121	136	108	139	-47	0.41	292	1.65	129	0.52
A2	Median	301	76	145	77	75	-66	0.25	209	0.72	34	0.16
	IQR	156	27	77	19	39	15	0.07	24	0.24	13	0.08
	Min	203	45	90	54	33	-99	0.12	154	0.43	16	0.09
	Max	888	213	481	127	271	-44	0.38	245	2.15	101	0.32
B1	Median	362	127	185	67	108	-68	0.33	224	0.72	71	0.33
	IQR	171	65	77	13	80	21	0.08	20	0.53	24	0.08
	Min	210	73	101	45	37	-92	0.16	167	0.42	40	0.24
	Max	1599	535	850	94	559	-42	0.40	258	3.58	238	0.40
B2	Median	859	280	606	78	189	-65	0.19	237	0.29	72	0.28
	IQR	314	94	135	19	148	12	0.09	17	0.09	37	0.05
	Min	210	73	101	45	37	-92	0.16	167	0.42	40	0.24
	Max	1782	653	1040	96	461	-39	0.34	264	0.63	159	0.36

Note: IQR = Interquartile range

Table 10 presents the results of the eye movement analyses per CEFR-level tasks. Differences can be observed in the eye-tracking measures between the four CEFR-level tasks, resulting from running Kruskal-Wallis analyses. The statistics expressed for each measure on CEFR task level are the median values across all participants. It can be seen that, with the exception of the measure, *Median length of regressions*, there are statistically significant differences for all the measures:

- *global processing measures*: 'total fixation time spend on text and response(s)' and 'total number of fixations'
- *text processing measures*: 'number of forward saccades', 'median length of forward saccades', 'number of regressions', 'median length of regressions', 'proportion of regressive movements', 'median fixation duration', 'sum fixation time on text per word'
- *item processing measures*: 'number of AOI switches between text and responses' and 'proportion of time spent fixating on responses'.

However, the *post hoc* pairwise comparisons show that statistically significant differences do not exist between all four CEFR-level item groups on each of the eye-tracking measures (see Table 10). For the global processing measures, differences were detected between all levels (with the exception of the number of fixations on A2 and B1 levels tasks). In terms of text processing, a mixed picture occurred with some measures showing differences especially between non-adjacent levels and other measures suggesting similar eye movement processes across CEFR level tasks (see Table 10). The item processing measures indicated differences between most CEFR task levels (but not B1-B2).

Table 10: Results eye-tracking analyses in relation to reading task CEFR level (RQ1a)

	Global processing measures		Text processing measures							Item processing measures	
	Total number of fixations	Total fixation time on text and response (seconds)	Number of forward saccades	Median length of forward saccades (px)	Number of regressions	Median length of regressions (px)	Proportion of regressive movements	Median fixation duration (ms)	Sum of fixation time on text per word (seconds)	Number of Aoi switches between text and responses	Proportion of time spent fixating on responses
Items											
A1 Items	188	49.7	55.5	70.75	54.5	-69	0.36	220	0.64	71	0.45
A2 Items	301	76.1	145	77	74.5	-66	0.25	209	0.72	34	0.16
B1 Items	362	127.2	185	66.61	108	-68.35	0.33	224	0.72	71	0.33
B2 Items	859	280	606	77.95	189	-64.77	0.19	237	0.29	72	0.28
Kruskal-Wallis											
H	69.04	79.6	79.11	11.97	49.04	3.99	46.47	18.65	54.19	35	79.92
df	3	3	3	3	3	3	3	3	3	3	3
P-Value	.000	.000	.000	.008	.000	.263	.000	.000	.000	.000	.000
Pairwise comparisons											
A1-A2	**	*	***				***			***	***
A1-B1	***	***	***		***						**
A1-B2	***	***	***		***		***	*	***		***
A2-B1		*					**			***	***
A2-B2	***	***	***		***			***	***	***	**
B1-B2	**	*	***	*			***	*	***		

For each of the eye-tracking measures, a hypothesis had been formulated on the direction of the measure and processing in relation to CEFR level (see Table 3). Table 11 restates these hypotheses and indicates whether support for these has been found in the statistical analyses presented in Table 10. A tick (✓) signifies that the hypothesis was fully supported; a cross (✗) means that support was not found; and both a tick and a cross (✓✗) signify that there was limited support for the hypothesis.

Table 11: Eye-tracking support for RQ1a hypotheses

	Measure	Hypothesis with reference to RQ1a	Met?
Global processing	Total number of fixations	As the CEFR level of the tasks increases, the total number of fixations will increase. This is directly due to the fact that the texts for the harder tasks are longer. Longer texts test higher-level cognitive processes which relate to comprehension at the sentence level and above.	✓
	Total fixation time on text and responses	As the CEFR level of the tasks increases, the total fixation time on the text and responses will increase. This measure is closely linked with 'total number of fixations', but it is more sensitive to the total time spent on the text and less sensitive to movement around the text.	✓
Text processing	Number of forward saccades	As the CEFR level of the tasks increases, the number of forward saccades will increase. This measure represents the fact that higher-level tasks have longer and more complex texts that require more processing.	✓
	Median length of forward saccades	As the CEFR level of the tasks increases, the median length of a forward saccade will decrease. This is due to the increased cognitive load on the test-taker as a function of CEFR level (higher level, higher cognitive processing load).	✓✗
	Number of regressions	As the CEFR level of the tasks increases, the total number of regressions will increase. This is due to two factors; firstly, a regression is more likely in a longer text, and secondly, regressions are more likely as the text becomes more challenging.	✓
	Median length of regressions	As the CEFR level of the tasks increases, the median length of the regressions will increase. This relates to the notion that more complex texts in the higher-level CEFR tasks will generate more between-word regressions, as they are designed to measure more higher-level cognitive processing, than the lower CEFR level tasks.	✗
	Proportion of regressive movements	As the CEFR level of the tasks increases, the proportion of regressive movements will decrease. As the texts increase in complexity, there will be a greater need to perform regressions to facilitate comprehension.	✗
	Median fixation duration	As the CEFR level of the tasks increases, the median fixation duration will increase. This would be due to the test-takers requiring longer fixations to comprehend the more complex texts and perform the more complex operations required by the higher-level tasks.	✓✗
	Sum fixation time on text per word	As the CEFR level of the tasks increases, so does the proportion of time spent on the text per word. This would be due to the fact that the increasing cognitive demands placed on the test-takers by the higher level texts require greater processing per word.	✓✗
Task processing	Number of Aoi switches between text and response options	As the CEFR level of the tasks increases, the number of switches between the text and the responses will increase. This would be due to the increasing difficulty in integrating the information contained in the text with the response in the selection of the correct answer.	✗
	Proportion of time spent fixating on response options	As the CEFR level of the tasks increases, the proportion of time spent fixating on the responses will decrease. This would be due to the proportionally increasing demand of the text over the responses as the level of the tasks increases.	✓✗

The support for the hypotheses on the global processing measures *Total fixation time on text and responses* and *Total number of fixations* show that the more complex tasks (i.e., higher CEFR-level target) elicited more processing overall from the participants. Similarly, the *Number of forward saccades* and *Number of regressions* on the text of the tasks increase as a function of task CEFR band. These results are intuitively logical, as one would expect longer texts to elicit more processing. However, it is interesting to note that there is a sharp increase in all these measures between the B1 tasks and the B2 tasks. This suggests a great increase in processed information in the B2 tasks over the other CEFR-level tasks.

Surprisingly, no support was found for the hypotheses relating to *Median length of forward saccades* and *Median length of regressions*. This is likely due to the fact that the individual variance in these measures is greater than the between item variance, and because of the small sample-size, potential effects could not be picked up. However, we can see a statistically significant difference between the B1 and B2 tasks on *Median length of forward saccades*. Potentially, differences in the goal of reading associated with these two CEFR level tasks explain this finding. As evidenced in the heat maps in Figure 5, the B1 tasks, which require filling in gaps in sentences selecting words from a bank, seem to have resulted in more careful local reading (i.e. around the gaps). On the other hand, the B2 tasks, which concern matching headings to the paragraphs of a long text, seemed to involve test-takers in more expeditious, global reading (i.e. for gist). Longer saccade lengths may be more efficient when the gist of a text is the main goal of comprehension and detailed comprehension is not required.

The results for *Proportion of regressive movements* do not support our hypothesis. However, an interesting pattern still emerges. The A1 and B1 texts generated a greater proportion of regressive movements than the A2 and B2 texts. The A1 and B1 tasks both concerned gap-fill formats, whereby the test-takers needed to decide which word of a set of given words best fit the gap. It is not unlikely that this required more local parsing (which is indeed visible in the heat maps) and (re)consideration of whether the words fit make sense grammatically, semantically, and are in line with the mental model of the text. This more local, careful (re)processing, therefore, was associated with proportionally more regressive movements.

There is limited support for the hypothesis regarding *Median fixation duration*. There is an increasing trend in fixation durations, likely due to the necessity to fixate longer on more complex texts. However, the A2 texts do not follow this trend. The *Sum of fixation time on text per word* gives a particularly interesting result, with regards to exactly how much of the test is being read in specific items. The post hoc analyses show a statistically significant difference between the B2 tasks and all other tasks. An examination of eye-trace plots provided a reason for this disparity, namely, many of the participants did not read the whole of the text if they felt they could correctly respond by just reading the first few lines. The result on the eye-tracking measure *Sum of fixation time on text per word* thus seems to suggest that the test-takers adopted a more expeditious or global sampling approach to the B2 tasks.

There was no support for the hypothesis regarding the item processing measure *Number of AoI switches between text and responses*; a similar level of switching was made for the A1-B1-B2 tasks.⁷ In terms of the *Proportion of time spent fixating on responses*, there are statically significant differences between all tasks, with the exception of B1-B2. However, these differences are seemingly not in line with the hypothesis set out above. The highest proportion was found for the A1 tasks, which is as expected. This may be related to the local parsing and consideration time needed for each of the three options provided per gap (and as compared to the shortest texts of all the texts in the different Aptis tasks). The deviating finding for the A2 tasks, however, is likely to be due to a technical difference in what constitutes the response AoI. The B1 and B2 tasks both involve a list of options to consider which remains the same for all items in the task. Presumably test-takers partly

⁷ It should be noted that the diverging figure for the A2 tasks is due to a task presentation difference. The AoIs for the responses to the A2 items constituted the section of the screen where the test-taker could select the number to signify the order of the sentences. This contrasts with the other CEFR level tasks where the response contained textual information in itself (e.g. words in the MC options, words in the bank, headings) and the response AoI thus carried text that needed to be comprehended, and carried meaning in itself. It is most likely that the low level of switching and fixating on responses in the A2 tasks is related to the fact that there was no textual information in the responses for the A2 items.

rely on memory or have to do less response processing when having processed the B1 and B2 response list a couple of times already.

5.2.2 Eye-tracking findings on cognitive processes depending on test-takers' L2 (reading) proficiency (RQ1b & RQ1c)

The second and third sub-questions of the study (RQ1b and c) were:

Are there any differences in cognitive processes depending on test-takers' L2 (reading) proficiency, as measured by the Aptis reading component?

To be able to answer these research questions, the test-takers' eye movements were analysed according to the test-takers' English reading proficiency (RQ1b) and according to their overall English language proficiency (RQ1c), as determined by their performances on the full Aptis test.

While a number of processing differences were found between the CEFR-linked tasks, the analysis of eye-movements as a function of test-taker ability uncovered fewer relationships. Only those measures which contained statistically significant relationships are reported here: *Total fixation time on text and responses*, *Proportion of time looking at the responses*, *Number of switches between the text and the responses*, and *Sum fixation time on text per word*. Table 12 shows the Spearman's correlation coefficients between four of the eye-tracking measures and the participant's score on the Aptis reading component (RQ1b) and all components of the Aptis (RQ1c). The statistically significant correlations at the 0.05 level have been highlighted in yellow in Table 12. Since the small sample size in this study may have masked some smaller yet extant effect sizes, the statistically significant correlations at the 0.1 level have also been highlighted (in green) for consideration.

Table 12: Results eye-tracker analyses in relation to L2 reading proficiency (RQ1b) and overall L2 proficiency(RQ1c)

			Total fixation time on text and responses	Proportion of time looking at the responses	Number of switches between the text and the responses	Sum fixation time on text per word
All items	L2 reading proficiency	Coefficient P-Value	-.387 .056	-.085 .685	-.319 .120	-.352 .084
	Overall L2 proficiency	Coefficient P-Value	-.411 .041	-.029 .890	-.385 .058	-.355 .081
A1 items	L2 reading proficiency	Coefficient P-Value	-.120 .568	-.139 .508	.002 .991	-.168 .423
	Overall L2 proficiency	Coefficient P-Value	-.150 .474	-.135 .518	-.166 .427	-.277 .180
A2 items	L2 reading proficiency	Coefficient P-Value	-.526 .007	-.149 .477	-.500 .011	-.497 .012
	Overall L2 proficiency	Coefficient P-Value	-.408 .043	.061 .773	-.288 .162	-.477 .025
B1 items	L2 reading proficiency	Coefficient P-Value	-.230 .268	.367 .071	-.232 .265	-.444 .026
	Overall L2 proficiency	Coefficient P-Value	-.202 .332	.235 .259	-.191 .360	-.351 .086
B2 items	L2 reading proficiency	Coefficient P-Value	-.269 .193	-.526 .007	-.394 .052	-.293 .155
	Overall L2 proficiency	Coefficient P-Value	-.329 .109	-.395 .050	-.399 .048	-.376 .064

Note: The statistically significant correlations at the 0.05 level have been highlighted in yellow. The statistically significant correlations at the 0.1 level have also been highlighted in green.

As can be seen in Table 12, all of the statistically significant results to the 0.05 level are in a negative direction, meaning that the measures diminished as the ability of the test-takers improved. Thematically, all of the measures might be interpreted as exemplifying different facets of the greater efficiency of processing by the better performing test-takers.

The results show that test-takers' L2 reading proficiency and overall L2 proficiency are negatively correlated with their total fixation time; the more proficient participants spent less time fixating on the reading tasks' texts and responses. This measure is also statistically significant for the A2 items, which suggests that the higher-level, between-sentence processing of the more proficient participants was greatly more efficient than that of the lower ability candidates, and enabled the former to respond more quickly. The fact that all the correlations for this measure are negative, even though only 3 of 10 are significant at the 0.05 level, suggests that this effect may be found in all items, but that the sample size used in this study did not provide sufficient power to detect it.

The results on *Proportion of time looking at the response options* for the B2 items show that the less proficient participants spent a greater proportion of their time looking at the responses than did the more proficient participants. This might relate to the efficiency of the higher ability participants at linking the mental model of the paragraphs to the correct response without the need for extensive consideration and reconsideration.

There are some statistically significant differences (at the 0.05 level) according to test-takers' L2 (reading) proficiency – and some approaching significance – related to the number of switches between text and response. All but one of the correlations is in a negative direction. These results are indicative of more efficient item resolution by the higher ability participants, and quite likely a larger sample size might find significance in all cases.

For the measure *Sum fixation time per word*, we can see that the more proficient participants spend less time, per word, processing the text. This is a manifestation of the increased processing efficiency of the higher ability students in terms of processing text more quickly and perhaps being more selective in what they read (in the case of the B2 items).

The interpretation of the above findings in relation to our hypothesis is presented in Table 13. A tick (✓) signifies that the hypothesis was fully supported; a cross (✗) means that support was not found; and both a tick and a cross (✓✗) signify that there was limited support for the hypothesis.

Overall, we can see that, on the basis of participants' eye movements, some processing differences have been found according to the test-takers' L2 (reading) proficiency (RQ1b and RQ1c), but that these were not overwhelming. This may be due to the relatively small sample size to detect statistical differences, and also the relatively high proficiency of the participants in the study in general (see Table 8). The in-depth analyses of the eye-tracking measures showed more evidence of differences in cognitive processing depending on the CEFR-linked task (RQ1a), whereby task type seemed to matter.

Table 13: Eye-tracking support for RQ1b and RQ1c hypotheses

	Measure	Hypothesis with reference to RQ1b & RQ1c	RQ1b met?	RQ1c met?
Global processing	Total number of fixations	As the ability of the test-taker increases, the number of fixations required to complete a task will decrease. This reflects the increased processing efficiency of higher ability test-takers who are able to process the text with fewer fixations, i.e. they have fewer breakdowns in comprehension leading to re-reading text, they use longer saccades (thus fewer fixations) to process text and/or they find the correct response quickly, and are confident in their selection, without the need for extensive searches or validation of their response.	x	x
	Total fixation time on text and responses	As the ability of the test-taker increases, the amount of time it takes fixating on a task will decrease. This reflects the increased processing efficiency of higher ability test-takers (see 'Total number of fixations').	✓ x	✓ x
Text processing	Number of forward saccades	As the ability of the test-taker increases, the number of forward saccades on the text of a task will decrease. This reflects the increased processing efficiency of higher ability test-takers (see 'Total number of fixations').	x	x
	Median length of forward saccades	As the ability of the test-taker increases, the median length of forward saccades on the text will increase. This is due to more skilful readers being able to process more information during each fixation.	x	x
	Number of regressions	As the ability of the test-takers increases, the number of regressions will decrease. This is because higher-ability test-takers need to solve fewer processing issues.	x	x
	Median length of regressions	As the ability of the test-takers increases, so will the length of regressions. This is because higher ability test-takers have fewer problems with word recognition and lexical access and thus perform fewer shorter regressions.	x	x
	Proportion of regressive movements	As the ability of the test-taker increases, the proportion of regressive movements will decrease. This would be due to the effect of poorer test-takers' need to re-read sections of the text to facilitate comprehension.	x	x
	Median fixation duration	As test-taker ability increases, the median fixation duration will decrease. This would be due to the better readers processing the information at each fixation faster than the poorer readers.	x	x
	Sum fixation time on text per word	As the ability of the test-takers increases, the sum fixation time per word will decrease. This reflects the ability of the higher-level test-takers to process the information contained in the text more quickly than the lower level test-takers.	✓ x	x
Task processing	Number of Aoi switches between text and response options	As the ability of the test-taker increases, the number of switches between the text and the responses will decrease. This would reflect the better test-takers being more able to process the text of the task and hold the representation in memory and not require extensive switching to the responses.	✓ x	✓ x
	Proportion of time spent fixating on response options	As the ability of the test-takers increases, the proportion of time spent fixating on the responses will increase. This would be due to the comprehension of the texts of the tasks presenting a proportionally smaller challenge to the better readers than to the poorer readers.	✓ x	✓ x

5.3 Stimulated recall

A second set of data to gain insights into test-takers' cognitive processing comprised stimulated recalls produced immediately after Aptis reading task.

5.3.1. Stimulated recall findings on cognitive processes during Aptis reading test completion (RQ1)

The overarching research question (RQ1) was:

What cognitive processes do test-takers employ during Aptis reading task completion?

Insights were gathered through stimulated recalls (with eye movement recordings as the stimulus) on two test versions of the Aptis reading component. This constituted a total of eight tasks (two per CEFR level) and 50 items, completed by 25 test-takers. In 81% of the cases (1014 items), the test-takers answered the items correctly, whereas 18% of the attempts were unsuccessful (236 items). Table 14 gives an overview of the cognitive processes used by the participants during reading test completion, as observed through the stimulated recalls. The data for correctly and incorrectly answered items are presented separately. In several cases, test-takers indicated they relied on more than one type of cognitive processing to arrive at an answer.

Unfortunately, the quantitative data cannot be illustrated with quotes from the stimulated recalls in this report because the Aptis reading tasks used in the study were live test versions at the time of writing and item content cannot be revealed. Nevertheless, when test-takers made more general comments on their processing, quotes supporting the findings are included in the text below.

Table 14: Stimulated recall results on cognitive processes during Aptis reading test completion (RQ1)

	Processes	Frequency Item correct (n=1014; 100%)		Frequency Item incorrect (n=236; 100%)	
		No.	%	No.	%
Goal setting	Careful reading – global	603	59%	111	47%
	Careful reading – local	505	50%	113	47%
	Expeditious reading – skimming	85	8%	10	4%
	Expeditious reading – search reading	54	5%	10	4%
	Expeditious reading – scanning	3	0.3%	0	0%
Central core	Creating intertextual representation	1	0.1%	0	0%
	Creating text level representation	191	19%	31	13%
	Building a mental model	297	29%	57	24%
	Inferencing	239	24%	95	40%
	Establishing propositional meaning	297	29%	68	28%
	Syntactic parsing	229	23%	41	17%
	Lexical access	289	29%	56	24%
	Word recognition	0	0%	0	0%
Additional	Creating paragraph level representation	145	14%	22	9%
	Background knowledge	17	2%	5	2%
	Collocation	98	10%	2	0.8%
	Pure guess	2	0.2%	10	4%

The results in Table 14 indicate that a wide range of processes were used during Aptis reading test completion. For correctly answered items, this included: the lower-level processes of lexical access (29%), syntactic parsing (23%), and establishing propositional meaning (29%); and the higher-level processes of inferencing (24%), building a mental model (29%), creating a paragraph level representation (14%), and creating a text level representation (19%). Overall, the most used processes were lexical access, establishing propositional meaning, and building a mental model, as evidenced in the stimulated recall data. No explicit evidence of the use of word recognition was found in the dataset; this is not surprising due to its automatized nature, which makes it difficult to observe through verbal protocol methods. In addition to the processes specified in Khalifa and Weir (2009), several instances were found in which the test-taker explicitly stated to have primarily or solely relied on collocational knowledge to complete the item. These cognitive processes were most often adopted when the test-takers were carefully reading the texts, often approaching the texts more globally (59%) and in half of the cases reading at a local level (50%). Expeditionary reading strategies such as skimming and search reading were also reported, but less frequently (8% and 5%, respectively). This was backed-up by the eye traces visible in the eye-movement recordings.

As shown in Table 14, similar patterns were found for those items that had not been answered correctly, but it is interesting that inferencing, relatively, had been used more often when the answer was incorrect. In addition, hardly any evidence for the use of collocational knowledge was found for the incorrect items.

So, for the Aptis reading component as a whole, it can be said that the items elicit the entire spectrum of processes specified in the central core of the Khalifa and Weir (2009) model, with the exception of word recognition (which is likely to be due to the research method), and intertextual representation.

In a limited number of instances, the test-takers indicated that they had determined the answer based on background knowledge, or they had simply guessed. Background knowledge was more often associated with correct answers, whereas guessing was associated with incorrect answers.

To gain a more in-depth understanding of the cognitive processing during Aptis reading test completion, three sub-analyses were conducted, as presented below, according to:

1. tasks' target CEFR level
2. test-takers' reading proficiency
3. test-takers' overall proficiency.

The sub-analyses mainly focus on the correctly answered items, since from a validation perspective these should reflect the intended cognitive processes. However, this does not imply that test-takers' processing when not getting the item right should be disregarded; such data can reveal useful insights into quality aspects of a test and individual items. In fact, insights from the incorrectly answered items will be drawn upon in the discussion of the sub-analyses. However, it was judged that the processing associated with the incorrectly answered items was less central to the specific aims of the present study.

5.3.2 Stimulated recall findings on cognitive processes when completing items targeting different CEFR levels (RQ1a)

The first sub-question (RQ1a) was:

Are there any differences in cognitive processes between items targeting different CEFR levels (and the associated task types): CEFR A1 target/ MC gap-fill; CEFR A2 target/sentence ordering; CEFR B1 target/banked gap-fill; and CEFR B2 target/matching headings?

To be able to answer this research question, the stimulated recall data on the correctly answered items were analysed per CEFR level of the tasks. For CEFR A1 level, this comprised data on two MC gap-fill tasks with 10 items, totalling 213 correct answers given by the 25 participants. The A2-level data set constituted two sentence ordering tasks with 12 items, resulting in 247 correctly answered items. Participants gave 279 correct answers on the two B1 banked gap-fill tasks (14 items), and 275 on the two B2 matching headings tasks (14 items). Table 15 shows the test-takers' cognitive processing per target CEFR level of the tasks, as observed in the stimulated recall data. The columns give an indication of the range and amount of use of the different processes per CEFR level tasks, i.e. what processes are used. The percentages allow for comparisons at row level between the four CEFR level tasks, i.e. whether there are differences depending on CEFR level tasks.

Table 15: Stimulated recall results on cognitive processes when correctly completing items targeting different CEFR levels (RQ1a)

	Processes	A1 items (n=213; 100%)		A2 items (n=247; 100%)		B1 items (n=279; 100%)		B2 items (n=275; 100%)	
		No.	%	No.	%	No.	%	No.	%
Goal setting	Careful reading – global	94	44%	264	107%	60	22%	185	67%
	Careful reading – local	169	79%	42	17%	248	89%	46	17%
	Expeditious reading – skimming	7	3%	9	4%	7	3%	62	23%
	Expeditious reading – search reading	0	0%	10	4%	0	0%	44	16%
	Expeditious reading – scanning	0	0%	1	0%	0	0%	2	0.7%
Central core	Creating intertextual representation	0	0%	0	0%	0	0%	1	0.3%
	Creating text level representation	32	15%	131	53%	7	3%	21	8%
	Building a mental model	83	39%	146	59%	47	17%	21	8%
	Inferencing	70	33%	33	13%	38	14%	98	36%
	Establishing propositional meaning	88	41%	12	5%	134	48%	63	23%
	Syntactic parsing	66	31%	49	20%	114	41%	0	0%
	Lexical access	44	21%	91	37%	56	20%	98	36%
	Word recognition	0	0%	0	0%	0	0%	0	0%
Additional	Creating paragraph level representation	0	0%	0	0%	0	0%	145	53%
	Background knowledge	0	0%	0	0%	17	6%	0	0%
	Collocation	40	19%	0	0%	58	21%	0	0%
	Pure guess	0	0%	0	0%	2	0.7%	0	0%

A1 items

To arrive at the correct answers on the *A1 items*, the test-takers had most often established propositional meaning (41% of the item answers), and also often made use of the other lower-level processes of syntactic parsing (31%) and lexical access (21%). In 19% of the cases, participants had relied on collocational knowledge, which they explained as “these words often go together” or “this combination just sounds right”. Higher-level processes had also helped many test-takers to arrive at the right answer: inferencing (33%), building a mental model (39%), and creating a text level presentation (15%). With regard to the last category, it needs to be kept in mind that the text constituted a six-sentence letter format and thus was short.

Often, the stimulated recalls evidenced the use of more than one strategy to arrive at an answer. For example, explicit syntactic parsing was several times followed by, or combined with, establishing propositional meaning or the use of collocational knowledge.

In several cases, test-takers first read through the entire text before solving the items. In these cases, most read through the whole text carefully although some quickly skimmed it (3%). For example, Participant 3 said: “I read from the beginning to the end of the paragraph to get an idea what the story was about”. This global careful reading approach was conducted to gain a general impression of the text, but sometimes also as part of solving items (see below) (totalling 44%). Mostly, however, test-takers did careful local reading (79%) to arrive at the right answer to the A1 items. Again, it should be noted that sometimes a sequence of more than one approach was used in the item completion process.

The stimulated recall data also led to interesting observations at the item level. For example, syntactic parsing particularly occurred with the first item in test version 6, whereas those who did not get this item correct were often looking for meaning differences and did not (or were not able) to make use of syntactic information. The latter was not a helpful strategy in the case of this item because all MC options were semantically possible, but not syntactically. It could be argued that the strong need of syntactic parsing without semantic clues shifts this item more towards language-in-use than reading comprehension. Collocational knowledge was particularly relied on by many test-takers to solve, for example, the third item in test version 6, or (in combination with syntactic parsing) the second item in test version 5. Intersentential meaning building – higher-level processing – specifically directed test-takers to the right answer of the first and third item of test version 5. This was also observed for the second item of test version 6, which most participants only managed to solve after having read the sentence of the third item and inferring on the basis of the meaning of that sentence. This last set of items was also more noticeably associated with global reading approaches.

With regard to the fifth item of test version 6, examples were observed in the stimulated recall data of participants arriving at each of the three MC options through a process of inferencing and logical reasoning which all seemed plausible justifications for different options to be considered correct.

A2 items

The A2 items, part of sentence ordering tasks, led many participants to build a mental model (59% of the A2 items) by considering the order of a series of events described in several individual sentences, or to create a complete text level representation (53%; with texts being seven sentences long). In addition to these higher-level reading processes, the use of lower-level processes such as lexical access (37%) and syntactic parsing processes (20%) was also witnessed in the data. These lower-level processes were almost always used in combination with a higher-level process in order to provide the right answer. In fact, it was found that such combinations were often crucial to determining the correct answer. For example, if test-takers principally made use of typical text structure knowledge, they ended up selecting the wrong final sentence for the sentence ordering task of test version 6. In contrast, test-takers who explicitly made use of syntactic processing in addition to text structure knowledge arrived at the correct sentence order for the final parts of the text.

The test-takers principally reported using a careful global reading approach with the A2 tasks, often first reading through the sentences to get a general idea before trying to re-order and re-read. For example, Participant 1 stated: "I read all the sentences and got an idea it was about [...]", and Participant 5 said that at the beginning: "I read from the first sentence to the last, trying to understand the story from these unordered sentences". Sometimes they first skimmed the text (4%). During task completion, the participants also typically used a careful global reading approach. This combination of pre-reading and item completion reading meant that the text was considered more than once in several cases, which explains the 107% for the goal setting process 'careful reading – global'. In some cases, the participants read more locally (17%), often specifically considering individual words or verb tenses to inform their more global reading, or they quickly searched for words such as logical connectors (4%).

B1 items

Although the correct answer to the B1 items had sometimes involved higher-level processing (building a mental model, 17%; inferencing, 14%), more often lower-level processing had been involved. Namely, based on the stimulated recall data, 48% of the correct answers were found at least partly by establishing propositional meaning, 41% by syntactic parsing, and 20% by lexical access processes.

The participants had also primarily applied a careful local reading approach (89%) to solving the B1 items. This aligns with the large use of propositional meaning establishment and was also visible in the within-sentence, local syntactic parsing to arrive at the correct answer. Participant 14, for example, stated: "I spent some minutes reading the text before and after the blanks". In general, participants often did not make use of extra-sentential information or what they had read in other sentences to determine the answer to an item; sentences stood on their own. This was, for example, very clear for the third item in test version 3, where test-takers combined syntactic information of the words surrounding the gap with propositional meaning establishment of the words following the gap (and sometimes linking knowledge of the world to this part of the sentence). This within-sentence focus was also very strongly present in the items (with the exception of one) of test version 6, whereby participants indicated to have importantly relied on local syntactic parsing (e.g. the third item) and collocational knowledge (e.g. the second and fourth item) – sometimes in combination with other sentence-level processing.

Instances of somewhat more careful global reading and text level representation were mostly associated with a few items only in the tasks (the first item in both test versions, which was often left until the end to complete). Some participants felt they needed to get a better overall picture of the text to determine or to feel more confident about the answer to these items. Some test-takers also first considered the words in the bank: "First, I looked through all the words in the box below" (Participant 7); "I started off by reading the first sentence of this passage (getting an idea of what it is about), then I went on to look through the words in the list, after which I started to complete each item" (Participant 13).

The sub-analyses of the B1 items also showed that the use of background knowledge (as found in the entire dataset) can be fully accounted for by one particular task (test version 6) and occurred mostly with the sixth item. This item concerns a fact about a famous person. Although some test-takers answered this item through a process of inferencing from the item co-text, others stated they simply knew the fact and had not needed the information that followed. In a few other instances, making use of knowledge about the world seemed to have assisted participants in a process of logical reasoning on some of the items in test version 6.

In a number of cases, it was observed that when test-takers mainly focussed on higher-level processing, and seemingly ignored or were unable to make use of syntactic information in the sentence, they were not able to determine the correct answer. In contrast, those who did answer the item correctly often explicitly combined higher-level processing with syntactic knowledge. This was, for example, the case for the first item of test version 6.

B2 items

For a considerable number of the correctly answered B2 items, which required matching headings, test-takers had made use of inferencing (36% of the items), lexical access (36%), and establishing propositional meaning (23%). In more cases, however, test-takers had created a paragraph level presentation (53%), although often in combination with one or more of the above three processes (29%).

Text level representations were less often observed (8%). The careful global reading (67%) that took place concentrated on the paragraph level, as did the use of the expeditious reading approach of skimming (23%). For example, Participant 24 reported: "I was reading paragraph by paragraph and immediately selecting the headings for each paragraph". Interestingly, some test-takers started the B2 tasks by reading the heading options (rather than the text), stating that they did this to get an idea of the text: "most of these items are right headings for different paragraphs, so by looking through them I will get an idea of the outline of this passage" (Participant 9); "I began by reading the headings given, trying to understand them and predict what the story would be about" (Participant 11). So, they seemed to use the headings as a way to gain a text level representation. Search reading (16%) was also often connected to the headings, whereby test-takers tried to link these to words (synonyms; phrases) in the text as a conscious strategy to determine the answers. Participant 3, for example, referred to this as "my matching-up technique".

The four CEFR level tasks

As can be seen in the individual CEFR-level task analyses above, each elicits a range of reading processes as defined by Khalifa and Weir (2009), and each showed a spread of lower- and higher-level processes. The balance, however, differed between the CEFR levels. Based on the stimulated recall findings, correct answering of A1 and B1 items had more often relied on lower-level processes, such as syntactic parsing and establishing propositional meaning, and also making use of collocational knowledge. The principal use of lower-level processes was most notable for the B1 items. These two CEFR level items most often involved a careful local reading approach (although careful global reading also occurred in many cases with the A1 tasks). Correctly completed A2 and B2 items, on the other hand, were proportionally more associated with higher-level processes, such as building a mental model and text/paragraph level representation.⁸ Lexical access, in addition, also appeared to often play a role in determining the correct answer of the A2 and B2 items. For these two groups of items, the test-takers had most often adopted a careful global reading approach. Expeditious forms of reading were also proportionally more reported for these CEFR level items, in particular for the B2 items.

It is important to keep in mind, however, that each CEFR level was associated with one particular task format. Thus, any differences may also be due to, or influenced by, task type, and it is indeed likely that the task formats partially explain the cognitive processing differences between the CEFR groups of items. The A1 and B1 tasks both concern a form of gap-filling. The A1 tasks require the test-taker to select, for each sentence, a word from three options provided for a gap in the sentence. The B1 tasks require the selection of a word from a bank of words provided for a collection of sentences with gaps. Although it has been argued that the rational deletion of words in a text gives item writers more control over what is being tested (Alderson, 2000), others have pointed out that there are no guarantees that the omitted words will lead to testing what is intended to be tested (Yamashita, 2003). One of the key controversies of cloze-type tasks (whether the traditional cloze with n^{th} -word deletions or gap-fill with targeted deletions) has been whether they can measure global or just local reading, and higher- or just lower-level processes. Empirical findings on cloze tests are conflicting, with some indicating that they can measure higher-level reading processes (e.g. Bachman 1982, 1985), but many others indicating that they are poor measures of such processes (e.g. Alderson 1979, 1980). Gao and Gu (2008), who looked into an item type similar to the Aptis B1 items, found by means of verbal protocols, that the test-takers indicated to have most often conducted within-sentence processing, at the clause level, to arrive at the answer.

⁸ In practice, the A2 texts were paragraph length. The B2 texts were much lengthier and consisted of eight paragraphs. Most processing of these longer texts focussed on paragraph representation rather than text representation.

Similar trends were observed by McCray, Alderson and Brunfaut (2012) for banked gap-fill tasks of the PTE Academic. Therefore, it is not unlikely that the task types used for the A1 and B1 items at least partly explain the tendency to rely on lower-level processes to arrive at the answer.

The above does not imply, however, that there is an absolute, direct relationship between task type and cognitive processes. In fact, as shown in the sub-analyses, differences were observed in the cognitive processes that were (primarily) used for individual items within one task.

The fact that higher-level processes were observed for all tasks might be (partly) related to the participants' level of ability in this study, with the majority being at B2 and C levels of proficiency. So for some items, participants may have made (more) use of higher-level processes, although the answer could have been arrived at through lower-level processing only. Potentially, the higher-level processing may have given test-takers more certainty that they had chosen the right answer, although we have no means to verify this in the dataset. Differences in processing according to participants' reading proficiency have, for example, been observed by Yamashita (2003) for banked-gap fill tasks. She found that less skilled readers expressed a larger emphasis on lower-level processing, such as making use of local syntactic information, whereas more skilled readers used a global reading approach and only tended to complement this with local, syntactic processes to confirm their answers. Thus, having more able readers in the present study might have overestimated the use of higher-level processing.

5.3.3 Stimulated recall findings on cognitive processes depending on test-takers' L2 (reading) proficiency (RQ1b & RQ1c)

A second set of sub-analyses aimed to look more closely into the potential impact of test-taker ability on the nature of the cognitive processing during reading test completion, as had been found in Yamashita (2003). Therefore, additional sub-questions had been formulated, one exploring differences according to reading proficiency (RQ1b) and one examining differences according to overall language proficiency (RQ1c). The measures for the test-takers' reading and overall proficiency were obtained through the administration of the Aptis test system.

Cognitive processes depending on test-takers' L2 reading proficiency

The second sub-question of the study (RQ1b) was:

Are there any differences in cognitive processes depending on test-takers' L2 reading proficiency, as measured by the Aptis reading component?

In order to answer this research question, the stimulated recall data on the correctly answered items were analysed per the CEFR reading proficiency level of the participants. It should be noted that the Aptis system reports proficiency data as CEFR level A1-A2-B1-B2, but that no distinction between C1 and C2 is made between the highest levels of performance. The C levels fall outside of the target proficiency range of Aptis and thus less precise, targeted measurements are made at these highest levels.

Based on the Aptis reading component results, four test-takers were at level B1, 10 at B2, and 11 were in the CEFR C range of reading proficiency. The cognitive processes during reading test completion, as demonstrated in the stimulated recall data, of the test-takers at each of these levels are presented in Table 16. The columns give an indication of the range and amount of use of the different processes by each individual reading proficiency group. The mean (M) cognitive process use per person allows for an easier proportion interpretation. In addition, to allow for comparisons at row level between the three reading proficiency groups, i.e. whether there are differences depending on test-takers' reading ability, the data of the B1 and B2 groups have been corrected for the proportion of items that had been accurately answered by each group of participants ($M_{\text{corrected B1}} n_{\text{correct}}=160$ out of 200; $M_{\text{corrected B2}} n_{\text{correct}}=396$ out of 500; $M_C n_{\text{correct}}=499$ out of 550). Because of the small numbers per group and the relatively small cognitive processing observation points per individual (which is related to the substantial demands and resources needed for eye-tracking and stimulated recall investigations), no comparative statistics were run.

Table 16: Stimulated recall results on cognitive processes of correct items depending on test-takers' L2 reading proficiency (RQ1b)

	Processes	B1 reading proficiency (n=4)			B2 reading proficiency (n=10)			C reading proficiency (n=11)	
		No.	M	M corrected	No.	M	M corrected	No.	M
Goal setting	Careful reading – global	95	23.75	24.7	230	23.00	24.15	278	25.27
	Careful reading – local	72	18.00	18.72	197	19.70	20.69	236	21.45
	Expeditious reading – skimming	17	4.25	4.42	30	3.00	3.15	38	3.45
	Expeditious reading – search reading	11	2.75	2.86	25	2.50	2.63	18	1.63
	Expeditious reading – scanning	1	0.25	0.26	1	0.10	0.11	1	0.09
Central core	Creating intertextual representation	0	0	0	1	0.10	0.11	0	0
	Creating text level representation	31	7.75	8.06	75	7.50	7.88	85	7.73
	Building a mental model	54	13.50	14.04	104	10.40	10.92	139	12.64
	Inferencing	39	9.75	10.14	102	10.20	10.71	98	8.91
	Establishing propositional meaning	37	9.25	9.62	123	12.30	12.92	137	12.45
	Syntactic parsing	22	5.50	5.72	92	9.20	9.66	115	10.45
	Lexical access	52	13.00	13.52	117	11.70	12.29	120	10.91
	Word recognition	0	0	0	0	0	0	0	0
Additional	Creating paragraph level representation	18	4.50	4.68	54	5.40	5.67	73	6.64
	Background knowledge	8	2.00	2.08	2	0.20	0.21	7	0.64
	Collocation	12	3.00	3.12	39	3.90	4.10	47	4.27
	Pure guess	0	0	0	1	0.10	0.11	1	0.09
Total		469	117.25	121.94	1193	119.3	125.27	1393	126.62

B1 readers

For those items they answered correctly, the B1 readers used the full range of central core processes that had been used by all test-takers, including lower- and higher-level processes. On average, they particularly often reported to have built a mental model (M=13.50) and do lexical processing (M=13.00) to establish the correct answer. Inferencing (M=9.75) and establishing propositional meaning (M=9.25) were also processes they had used several times. They indicated to have employed these processes mainly while doing careful global reading (M=23.75) and/or careful local reading (M=18.00). In addition, they occasionally skimmed the text (M=4.25) or did search reading (M=2.75).

B2 readers

The B2 readers also made use of a wide range of lower- and higher-level cognitive processes when determining the correct answer to the various Aptis reading tasks used in this study. On average, they had most often indicated to establish propositional meaning ($M=12.30$), do lexical processing ($M=11.70$), build a mental model ($M=10.40$), infer ($M=10.20$), and to do syntactic parsing ($M=9.20$) as part of the process leading to the right answer. They most often read the texts carefully (global reading $M=23.00$; local reading $M=19.70$), and sometimes conducted expeditious reading (skimming $M=3.00$; search reading $M=2.50$).

C readers

The stimulated recall data showed that the C readers similarly employed a wide spectrum of lower- and higher-level cognitive processes during Aptis reading task completion. On average, they had relied more frequently on the processes of building a mental model ($M=12.64$), establishing propositional meaning ($M=12.45$), lexical access ($M=10.91$), and syntactic parsing ($M=10.45$). Like the other groups, they primarily read the texts carefully – with a more global ($M=25.27$) or local ($M=21.45$) comprehension goal. In some cases, they adopted skimming approaches ($M=3.45$).

The three reading proficiency levels

Although all three reading proficiency groups had used a wide variety of cognitive processes, some differences can be inferred from the stimulated recall data in terms of the extent to which test-takers of different reading ability had used particular types of processing when determining the correct answers. The data showed that, overall, more proficient readers on average reported a somewhat larger amount of cognitive processing while correctly solving the items ($M_{\text{corrected}_{B1}}=121.94$, $M_{\text{corrected}_{B2}}=125.27$, $M_{\text{corrected}_C}=126.62$). Since the stimulated recalls were conducted in the test-takers' L1 (and thus, most likely, they were easily able to formulate their thoughts out loud), and the eye-tracking traces served as a reminder of their processing (while also informing the research assistant's stimulated recall questions), it is rather unlikely that this overall trend results from the use of the stimulated recall method.

Several trends emerged for the lower-level processes (as defined in Khalifa & Weir, 2009). On average, less proficient readers had focussed more often on lexical processing ($M_{\text{corrected}_{B1}}=13.52$, $M_{\text{corrected}_{B2}}=12.29$, $M_C=10.91$), less often on syntactic parsing ($M_{\text{corrected}_{B1}}=5.72$, $M_{\text{corrected}_{B2}}=9.66$, $M_C=10.45$), and made less use of collocational knowledge ($M_{\text{corrected}_{B1}}=3.12$, $M_{\text{corrected}_{B2}}=4.10$, $M_C=4.27$). When looking into the dataset from the combined perspective of test-taker CEFR reading proficiency and CEFR-level task, it was found that the larger average use of syntactic processes and smaller average use of lexical processes with increasing test-taker reading proficiency was particularly associated with the A1 and B1 items (the two gap-fill task types).

In terms of the higher-level processes, more proficient readers tended to make more use of the process 'creating a paragraph level representation', a process that was associated with the B2 level tasks ($M_{\text{corrected}_{B1}}=4.68$, $M_{\text{corrected}_{B2}}=5.67$, $M_C=6.64$). The B readers, on the other hand had used inferencing slightly more often on average than the C readers as part of their successful item completion processes ($M_{\text{corrected}_{B1}}=10.14$, $M_{\text{corrected}_{B2}}=10.71$, $M_C=8.91$).

All groups had primarily carefully read the texts (global and local reading), but the less proficient readers had more often also employed expeditious reading approaches (skimming and search reading) than the more proficient readers.

Although the processing associated with incorrectly answered items is not presented in detail in this report, some observations from those data are interesting. Overall, the C readers on average reported comparatively slightly fewer cognitive processing than the other proficiency groups for those items they did not answer correctly ($M_{\text{corrected}_{B1}}=14.88$, $M_{\text{corrected}_{B2}}=14.18$, $M_C=12.91$). As part of the processing leading to incorrect answers, the test-takers most often conducted inferencing processes, and also more so if they were lower in reading ability ($M_{\text{corrected}_{B1}}=4.89$, $M_{\text{corrected}_{B2}}=3.48$, $M_C=2.64$). So, while we do not wish to overemphasize this finding, it may be suggested that inferencing is associated with a higher risk of making wrong deductions based on the information in the text, particularly for those with lower levels of reading ability.

Other tendencies found in the data for the incorrect answers are that less able readers' syntactic parsing on average was associated more often with incorrect answers than the average use of this process by more able readers ($M_{\text{corrected}_{B1}}=1.91$, $M_{\text{corrected}_{B2}}=1.70$, $M_C=1.00$), and less able readers had relied just slightly more on guessing ($M_{\text{corrected}_{B1}}=0.64$, $M_{\text{corrected}_{B2}}=0.32$, $M_C=0.27$). However, due to the low number of observations of these processes, these findings need to be treated with great care.

In sum, B1-C readers all make use of a wide range of processing, but some trends are noticeable in terms of which processes are used more/less on average. It should be kept in mind, however, that the number of participants was relatively small, in particular for the B1 group. Potentially, though, the trends would be more prominent with a larger sample and/or more 'weaker' L2 readers.

Cognitive processes depending on test-takers' overall L2 proficiency

The third sub-question of the study (RQ1c) was:

Are there any differences in cognitive processes depending on test-takers' overall L2 proficiency, as measured by all different Aptis components?

To answer this research question, the stimulated recall data on the correctly answered items were analysed for two groups of test-takers. The groupings were based on participants' total scores on all four Aptis skill components (listening, reading, writing, speaking),⁹ i.e. the 12 lowest scoring participants formed the 'lower proficiency half' ($M=155.5$, $SD=8.6$), and the 13 highest scoring participants constituted the 'higher proficiency half' ($M=177$, $SD=4.7$). It should be noted that the Aptis system does not report CEFR levels for the overall performance on all components.

The two proficiency groups' cognitive processes during reading test completion, as demonstrated in the stimulated recall data, are presented in Table 17. The columns give an indication of the range and amount of use of the different processes by each proficiency half. The mean (M) cognitive process use per person allows for an easier proportion interpretation. In addition, to allow for comparisons at row level between the two groups, i.e. whether there are differences depending on test-takers' overall English language proficiency, the data of the lower proficiency group have been corrected for the proportion of items that had been accurately answered by each group of participants ($M_{\text{corrected}_L} n_{\text{correct}}=466$ out of 600; $M_H n_{\text{correct}}=548$ out of 650). Because of the relatively small numbers per group and the small number of cognitive processing observation points per individual (which is related to the substantial demands and resources needed for eye-tracking and stimulated recall investigations), no comparative statistics were run.

⁹ The 'All components score' reported by the Aptis system is the sum of test-takers' scores on each of the four skills tests.

Table 17: Stimulated recall results on cognitive processes of correct items depending on test-takers' L2 proficiency (RQ1c)

	Processes	Lower proficiency half (n=12)			Higher proficiency half (n=13)	
		No.	M	M corrected	No.	M
Goal setting	Careful reading – global	274	22.80	24.62	329	25.31
	Careful reading – local	232	19.33	20.88	273	21.00
	Expeditious reading – skimming	34	2.83	3.06	51	3.92
	Expeditious reading – search reading	29	2.42	2.61	25	1.92
	Expeditious reading – scanning	2	0.17	0.18	1	0.08
Central core	Creating intertextual representation	1	0.08	0.09	0	0
	Creating text level representation	89	7.42	8.01	102	7.85
	Building a mental model	129	10.75	11.61	168	12.92
	Inferencing	113	9.42	10.17	126	9.69
	Establishing propositional meaning	146	12.17	13.14	151	11.62
	Syntactic parsing	93	7.75	8.37	136	10.46
	Lexical access	147	12.25	13.23	142	10.92
	Word recognition	0	0	0	0	0
Additional	Creating paragraph representation	53	4.42	4.77	92	7.08
	Background knowledge	9	0.75	0.81	8	0.62
	Collocation	39	3.25	3.51	59	4.54
	Pure guess	0	0	0	2	0.15
Total		1390	115.81	125.07	1665	128.08

Lower proficiency half

Participants in the lower proficiency group reported using a wide range of lower- and higher-level cognitive processes to successfully complete Aptis reading items. On average, they specifically often used the lower-level processes 'lexical access' (M=12.25) and 'establishing propositional meaning' (M=12.17), as well as the higher-level processes 'building a mental model' (M=10.75) and 'inferencing' (M=9.42). Most often they carefully read the texts, focussing on a more global (M=22.80) or local (M=19.33) picture, although they also sometimes used expeditious forms of reading.

Higher proficiency half

The higher proficiency group similarly adopted various lower- and higher-level cognitive processes while establishing the correct answers. Particularly, they often built a mental model (M=12.92), established propositional meaning (M=11.62), did lexical (M=10.92) and syntactic (M=10.46) processing, and inferred (M=9.69). Most often, they carefully read the texts, focussing on a more global (M=25.31) or local (M=21.00) picture, although they also sometimes used expeditious forms of reading, such as skimming (M=3.92).

The two proficiency groups

Although both proficiency groups used a wide variety of cognitive processes, the stimulated recall data suggest a few different tendencies in the extent to which test-takers of different L2 proficiency use particular types of processing when determining the correct answers.

The most notable differences between the two groups concerned the higher average use of syntactic parsing processes ($M_{\text{corrected}_L}=8.37$, $M_{\text{corrected}_H}=10.46$) and paragraph representation processes ($M_{\text{corrected}_L}=4.77$, $M_H=7.08$) by the more proficient test-takers. The latter is all accounted for by the B2 items. The syntactic parsing differences between the two groups was most prominent with the A1 and B1 gap-fill tasks. On the other hand, less proficient test-takers on average used comparatively more lexical processing ($M_{\text{corrected}_L}=13.23$, $M_H=10.92$) and propositional meaning building processes ($M_{\text{corrected}_L}=13.14$, $M_H=11.62$).

Somewhat less pronounced trends are the slightly higher average reliance on collocational knowledge by more proficient test-takers ($M_{\text{corrected}_L}=3.51$, $M_H=4.54$), and building a mental model ($M_{\text{corrected}_L}=11.61$, $M_H=12.92$). The collocation use differences between the two groups were specifically associated with the B1 banked-gap fill tasks. The group differences in reliance on 'building a mental model' occurred mostly with the A1 tasks.

Although the processing associated with incorrectly answered items is not presented in detail in this report, one observation from those data is interesting to mention. Overall, the type of processing conducted by both proficiency groups was very similar, but inferencing was slightly more often involved in the cognitive processing of items which the less proficient test-takers answered wrongly ($M_{\text{corrected}_L}=3.56$, $M_H=2.62$), as well as syntactic parsing ($M_{\text{corrected}_L}=1.69$, $M_H=0.92$).

All in all, the differences according to proficiency groups are relatively limited, with the exception of some tendencies. Potentially, more obvious differences would have been found if the population of the study had been more diverse in terms of L2 proficiency.

6. DISCUSSION

The above findings show that the Aptis reading component elicits a wide range of cognitive processes from test-takers while they are successfully completing the items. With the exception of intertextual representation, a multitude of examples of the various processes defined by Khalifa and Weir (2009) were observed in the dataset. The Aptis reading component as a whole was thus found to sample extensively from the construct of reading in terms of cognitive processing.

In its Aptis Candidate Guide, the British Council (2013) gives more specific descriptions of the nature of the four different parts of the Aptis reading component, thereby also including information related to cognitive processes (see Figure 6).

For the A1 items, the Candidate Guide (British Council, 2013, p. 11) specifies that "[e]ach sentence in the short text is free-standing but appears to form a text, so it is not necessary to understand all of the sentences to answer individual questions". The Candidate Guide also advises test-takers that this task "assesses your ability to read a sentence and to complete it with an appropriate grammatical form or word" (p. 11). Both the eye-tracking and the stimulated recall analyses confirmed that the successfully completed A1 items had engaged the test-takers in an important amount of sentence-level processing, such as lexical access, syntactic parsing, and propositional meaning building, and that they very often used a careful local reading approach. Sometimes, however, test-takers also used a more global, extra-sentential reading approach and employed higher-level processes to correctly complete the A1 items (whether or not in combination with local, lower-level processing). In fact, this had appeared to be crucial in the case of a few items (see section 5.3.2). Thus, overall evidence was found for the intended processing of A1 items, but a few individual items appeared to require slightly different, higher-level types of reading processes.

Figure 6: Aptis reading test overview for candidates (British Council, 2013, p. 11)

Test design	Description	Preparation
Part 1 Sentence comprehension	Choose drop-down words to complete sentences. Each sentence in the short text is free-standing but appears to form a text, so it is not necessary to understand all of the sentences to answer the individual questions. There are five multiple choice questions in total, each with three options.	This part of the reading test is aimed at CEFR level A1 (the lowest) and assesses your ability to read a sentence and to complete it with an appropriate grammatical form or word. To prepare for this task, it would be useful to go back to the grammar and vocabulary activities mentioned in the description of the core test above. Of course the best way to become a better reader is to practise. A number of publishers produce graded readers that might be of use. For example, try: Cambridge Bookworms Starter/Stage 1. Cambridge Readers – Level 1. Penguin Readers – Level 1. Macmillan Readers – Starter/Beginner. Headway Skills series.
Part 2 Text cohesion	In this task you will see a series of seven sentences. They belong to a single story that has been jumbled up. There is only one way that the sentences go together to form the story and your task is to click on the sentences and drag them to the correct position in the story. This task tests your knowledge of the cohesion of a text. So, you are looking for the clues in each sentence that show how it links to other sentences.	Read all of the sentences carefully first. Then, decide on the order (the first sentence is identified for you). Appropriate readers for this level are: Cambridge Bookworms Stage 1 and 2. Cambridge Readers – Level 2. Penguin Readers – Level 3. Macmillan Readers – Elementary.
Part 3 Short-text comprehension	In this task you will need to read a short text (about 150 words). The task is to complete the text by selecting the appropriate words (from a list) to fill in the gaps. To complete all of the text you need to understand more than just a sentence.	Read over the whole text before attempting the questions. Appropriate readers for this level are: Cambridge Bookworms Stage 2 and 3. Cambridge Readers – Level 3, 4 and 5. Penguin Readers – Level 4. Macmillan Readers – Pre Intermediate.
Part 4 Long-text comprehension	This task consists of a long text (about 750 words) with a series of headings. The task is to match the headings to paragraphs in the text (there are seven to be done). There is always an extra heading that does not fit with any paragraph. This task is designed to test your ability to read and understand a long text. In addition, you need to be able to demonstrate an understanding of how the headings reflect the paragraphs in different ways (sometimes using similar words, sometimes similar ideas, or by sharing a topic – though this is never obvious).	Read the main text carefully but as quickly as you can. Then carefully read the headings. Do all this before starting the task. Look for clues to connect the headings to the paragraphs; these might be similar words, ideas or topics. Appropriate readers for this level are: Cambridge Bookworms Stage 4, 5 and 6. Cambridge Readers – Level 4, 5 and 6. Penguin Readers – Level 5 and 6. Macmillan Readers – Intermediate and Upper Intermediate.

As can be seen in Figure 6, with the second task type the test developers aimed to assess “knowledge of the cohesion of a text. So, you are looking for the clues in each sentence that show how it links to other sentences” (British Council, 2013, p. 11). The test developers also recommend that the test-takers “[r]ead all of the sentences carefully first” (p. 11). This description suggests a careful global reading approach with specific attention to words such as logical connectors. Evidence for this type of processing was indeed found in the dataset. For example, the heat map visualisations showed that test-takers paid particular attention to the temporal adjectives, anaphors and logical connectors at the beginning of the sentences, and the stimulated recalls showed lexical processing of these words. At the same time, successful test-takers had often used a careful global reading approach to read through all the sentences and built a mental model and/or made text level representations to complete these items. Some test-takers had skimmed the text first rather than starting off with careful reading.

On the third task type, the Candidate Guide (p. 11) states that “[t]o complete all of the text you need to understand more than just one sentence”, and it recommends “[r]ead[ing] over the whole text before attempting the questions”. The eye-tracking and stimulated recall findings on this part of the test included evidence of some higher-level processing and careful global reading and skimming approaches. Such processing appeared often crucial for some specific items. For example, several test-takers only completed the first gap at the end, after having read through and completed all other gaps. However, the majority of cognitive processes associated with the successfully completed items of this task type showed to have involved careful local reading and sentence-level processing. Test-takers made use only of information within the sentence. Thus, although this part of the test measures reading processes, they often seem to be different from the processes the test developers intended to be measured with this specific task.

Also, overall, test-takers had to establish an understanding of what they had read to solve the items and consider various options in the banks of words to determine the answer. However, it was observed that some test-takers completed one specific item primarily on the basis of background knowledge (but it should be said that many others inferred on the basis of within-sentence reading). Since this task type makes use of ‘familiar’ text topics, there is a risk of construct-irrelevant variance, which seemed to have occurred with one particular item for a few test-takers. However, it was not an overall task issue, and very careful, explicit item writer guidance may help avoid individual items that target a well-known fact.

Another observation of the third task type, which was also made for the first task type (both gap-fill tasks), is that some items were successfully solved by several test-takers on the basis of collocational or syntactic knowledge (sometimes in combination with other types of processing, sometimes less so). Depending on the extent of reliance on this type of knowledge to successfully complete the item, this may constitute construct-irrelevant variance and test language-in-use more so than reading comprehension. The type of distractors provided seemed to play a role in the extent to which collocational or syntactic knowledge primarily or solely determined the answer. Again, this was not an overall task-level risk, but associated with particular items. Specific item writer guidance may reduce this potential threat of construct-irrelevant variance.

Part 4 of the Aptis reading component “is designed to test your ability to read and understand a long text. In addition, you need to be able to demonstrate an understanding of how the headings reflect the paragraphs in different ways (sometimes using similar words, sometimes similar ideas, or by sharing a topic – though this is never obvious)” (British Council, 2013, p. 11). The advice given to candidates is to “[r]ead the main text carefully but as quickly as you can. Then carefully read the headings. Do all this before starting the task. Look for clues to connect the headings to the paragraphs; these might be similar words, ideas or topics” (p. 11). These types of processing were observed in the eye-tracking and stimulated recall data which indicated a more global careful reading approach, with sometimes expeditious reading and search reading for words/ideas to match to the headings. One noticeable deviation is that the test-takers in the study did not start off by reading the entire text; in practice, they adopted a more paragraph-level global reading approach (often after having inspected the list of headings first).

The above findings and discussion provide key information on what the Aptis reading component tests (as observed during Aptis reading task completion of the participants in this study), and how this compares with what is intended to be tested. Thus, the study provides vital information for Aptis test validation. A summary of the key findings in this respect is provided in the section below.

7. CONCLUSION

The primary aim of this study was to examine test-takers' cognitive processing while responding to Aptis reading comprehension items. Through a process of eye-tracking during task completion, followed by stimulated recalls, rich insights were gained in the cognitive processes employed during reading item completion.

The innovative methodology – including a detailed analysis of eye movement metrics and of the stimulated recalls test-takers produced – proved to be particularly useful. Eye-tracking visualisations revealed several overall processing patterns and tendencies, which were triangulated by the stimulated recall findings. Although both the eye-tracking and stimulated recall analyses led to data on both lower- and higher-level processing, the eye movement analyses allowed for relatively more insights into lower-level reading processes, whereas the stimulated recall data were useful in revealing relatively more higher-level reading processes. In addition, the findings resulting from the two methods mutually confirmed the other results, thus providing a solid basis on which to draw conclusions on test-takers' cognitive processes. As a consequence, the researchers believe that this methodology could be extremely valuable as part of test validation research, even though it is quite labour intensive. For example, it could be applied to provide valuable empirical *a priori* validation evidence based not on “what the test constructors believe an item to be testing”, but on what processes underlie correct item responses (Alderson, 2000, p. 97). As a result, it could, for example, help test constructors deploy batteries of items into their tests which more accurately reflect the overall test construct, helping to minimise the two major threats to validity: construct under-representation and construct-irrelevant variance (Messick, 1992).

In this study, the use of the two methods showed that the entire range of cognitive processes as specified by Khalifa and Weir (2009) (with the exception of intertextual representation) was used by test-takers while completing the Aptis reading component. This suggests that the Aptis reading component, as a whole, quite comprehensively taps into the construct of reading. There was evidence of expeditious reading approaches, but the majority of correct items had involved test-takers in careful reading at a global and/or local level. The test-takers also engaged in lower-level processing (e.g. lexical access, syntactic parsing, propositional meaning building) as well as higher-level processing (e.g. inferencing, building a mental model, creating paragraph/text level representations). Risks of construct-irrelevant variance (e.g. the reliance on guessing and background knowledge for correctly answered items) were found only to a limited extent, and appeared to be associated with specific individual items rather than tasks or the test as a whole. Some test-takers' seemingly primary use of collocational knowledge (particularly more proficient participants) to complete a couple of gap-fill items might require further analysis to ensure a main focus on testing reading comprehension as opposed to language in use.

Notable differences were observed in the relative reliance on specific types of processing between items targeting different CEFR levels. However, these differences appeared to be associated with task type more so than with the target CEFR level of the tasks; gap-fill items elicited relatively more careful local reading and lower-level processing, while sentence-ordering and matching headings items involved relatively more careful global reading, higher-level processing, and some expeditious reading. With the exception of the B1 tasks, these patterns seem to be generally in line with the Aptis intended target processes for each CEFR-linked task level.

Some processing trends associated with test-takers' L2 (reading) proficiency were noticed, such as relatively more frequent use of syntactic parsing, collocations and paragraph level representations, but less frequent use of lexical access processes by more proficient L2 readers/language learners. However, these tendencies were not as outspoken as the differences associated with the tasks, which may be due to the overall relatively high proficiency of the participants in the study.

BIBLIOGRAPHY

- Alderson, J. C. (1979). The cloze procedure and proficiency in English as a Foreign Language. *TESOL Quarterly*, 30, 59–76.
- Alderson, J. C. (1980). Native and non-native speaker performance on cloze tests. *Language Learning*, 30, 219–223.
- Alderson, J.C. (2000). *Assessing reading*. Cambridge: CUP.
- Alderson, J.C., (1993). Judgments in language testing. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing* (pp. 46–53). Alexandria, VA, USA: TESOL.
- Alderson, J.C., Brunfaut, T., McCray, G., & Nieminen, L. (2012). *Component-skills approach to L2 reading: Findings, challenges, and innovations*. Paper presented at the Conference of the American Association for Applied Linguistics (AAAL), Boston (USA).
- Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing*, 8(1), 41–66.
- Andrews, S., Miller, B., & Rayner, K. (2004). Eye movements and morphological segmentation of compound words: There is a mouse in mousetrap. *European Journal of Cognitive Psychology*, 16, 258–311.
- Ashby, K., Clifton, C., & Rayner, J. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *Quarterly Journal of Experimental Psychology*, 17, 364–388.
- Bachman, L. F. (1982). The trait structure of cloze test scores. *TESOL Quarterly*, 16, 61–70.
- Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, 19, 553–556.
- Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing*, 28(1), 51–75.
- Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, 30(4), 441–465.
- Bax, S., & Weir, C. J. (2012). Investigating learners' cognitive processes during a computer-based CAE reading text. *Cambridge ESOL: Research Notes*, 3–14.
- Blanchard, H., Pollatsek, A., & Rayner, K. (1989). The acquisition of parafoveal word information in reading. *Perception and Psychophysics*, 46, 85–94.
- British Council (2013). *Aptis candidate guide – Online version*. Retrieved from <http://www.britishcouncil.org/aptis>
- British Council (2014). *Aptis*. Retrieved from <http://www.britishcouncil.org/aptis>
- Brunfaut, T., & McCray, G. (2014). *Looking into reading: Pilot study*. Unpublished report. British Council Assessment Research Grant 2012.
- Brysbaert, M., & Vitu, F. (1998). Word skipping: Implications for theories of eye movement control in reading. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 125–148). Oxford: Elsevier.
- Ceder, A. (1977). Drivers' eye-movements as related to attention in simulated traffic flow conditions. *Human Factors*, 19(6), 571–581.
- Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R. G. P. van Gompel, W. Murray, R. L. Hill, & M. H. Fischer (Eds.), *Eye movements: A window on the mind and brain* (pp. 341–371). Amsterdam: Elsevier.
- Council of Europe (2001). *Common European Framework of Reference for languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
- Dunlea, J. (2014, June). *Investigating the relationship between empirical task difficulty, textual features and CEFR levels*. Paper presented at the EALTA conference, University of Warwick, UK.

- Dussias, P. (2003). Syntactic ambiguity resolution in L2 learners: Some effects of bilinguality on L1 and L2 processing strategies. *Studies in Second Language Acquisition*, 25, 529–557.
- Dussias, P. (2010). Uses of eye-tracking data in second language sentence processing research. *Annual Review of Applied Linguistics*, 30, 149–166.
- Dussias, P., & Sagrara, N. (2007). The effect of exposure on syntactic parsing in Spanish–English bilinguals. *Bilingualism: Language and Cognition*, 10, 101–116.
- Ehrlich, S., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behaviour*, 20, 641–655.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.
- Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London: Sage.
- Frenck-Mestre, C. (2002). An online look at sentence processing in the second language. In R. Heredia & J. Altarriba (Eds.), *Bilingual sentence processing* (pp. 218–236). Amsterdam: Elsevier.
- Gao, X., & Gu, X. (2008). An introspective study on test-taking process of banked cloze. *CELEA Journal*, 31(4), 3–16.
- Gass, S., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Gorin, J. (2006, April). *Using alternative data sources to inform item difficulty modelling*. Paper presented at the Annual Meeting of the National Council on Educational Measurement, San Francisco, CA.
- Holmqvist, K., Nyström, M., Anderson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye-tracking*. Oxford: Oxford University Press.
- Hyönä, J., & Pollatsek, A. (1998). Reading Finnish compound words: Eye fixations are affected by component morphemes. *Human Perception and Performance*, 24, 1612–1627.
- Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception and Psychophysics*, 40, 431–439.
- Jacobson, J. Z., & Dodwell, P. C. (1979). Saccadic eye movements during reading. *Brain and Language*, 8(3), 303–341.
- Keating, G. (2009). Sensitivity to violation of gender agreement in native and non-native Spanish. *Language Learning*, 59, 503–535.
- Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge: Cambridge University Press.
- Levy-Schoen, A. (1981). Flexible and/or rigid control of oculomotor scanning behaviour. In D. F. Fischer, R. A. Monty, & J. W. Senders (Eds.), *Eye movements: Cognition and visual perception* (pp. 299–316). Hillsdale, NY: Erlbaum.
- May, J. G., Kennedy, R. S., Williams, M. C., Dunlap, W. P., & Brannan, J. R. (1990). Eye movement indices of mental workload. *Acta Psychologica*, 75(1), 75–89.
- McCray, G. (2013). *Statistical modelling of cognitive processing in reading comprehension in the context of language testing*. Unpublished PhD thesis. Lancaster University, UK.
- McCray, G., Alderson, J. C., & Brunfaut, T. (2012, June). *Validity in reading comprehension items: triangulation of eye-tracking and stimulated recall data*. Paper presented at the EALTA conference, University of Innsbruck, Austria.
- Messick, S. (1992). Validity of test interpretation and use. In M. C. Alkin (Ed.), *Encyclopedia of Educational Research* (6th ed.) (pp. 1487–1495). NY: Macmillan.
- Murray, W. S., & Kennedy, A. (1988). Spatial coding in the processing of anaphor by good and poor readers. *The Quarterly Journal of Experimental Psychology*, 40(4), 693–718.
- O'Sullivan, B. (2011). Language testing. In J. Simpson (Ed.), *Routledge handbook of applied linguistics*. Oxford: Routledge.
- O'Sullivan, B. (2012). Aptis test development approach. *Aptis technical report ATR-1*. British Council. <http://www.britishcouncil.org/aptis>

- O'Sullivan, B., & Weir, C.J. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing theories and practices* (pp. 13–32). Basingstoke: Palgrave Macmillan.
- Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language testing*, 20(1), 26–56.
- Pollatsek, A., Hyönä, J., & Bertram, R. (2000). The role of morphological constituents in reading Finnish compound words. *Journal of Experimental Psychology: Human Perception and Performance*, 26, 820–833.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372–422.
- Rayner, K., & Pollatsek, A. (1989). *The psychology of reading*. Englewood Cliffs, NJ: Prentice Hall.
- Rayner, K., & Well, A. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin and Review*, 3, 504–509.
- Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the E-Z reader model. *Journal of Experimental Psychology: Human Perception and Performance*, 30, 720–732.
- Rayner, K., Juhasz, B., & Pollatsek, A. (2007). Eye movements during reading. In M. Snowling & C. Hulme (Eds.), *The science of reading* (pp. 79–97). Malden, MA: Blackwell.
- Rayner, K., Pollatsek, A., Ashby, J., & Clifton, C. (2012). *Psychology of reading* (2nd ed.). New York: Psychology Press.
- Rayner, K., Warren, T., Juhasz, B., & Liversedge, S. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30, 1290–1301.
- Recarte, M. A., & Nunes, L. M. (2003). Mental workload while driving: Effects on visual search, discrimination and decision making. *Journal of Experimental Psychology: Applied*, 9(2), 119–137.
- Roberts, L., Gullberg, M., & Indefrey, P. (2008). Online pronoun resolution in L2 discourse: L1 influence and general learner effects. *Studies in Second Language Acquisition*, 30, 333–375.
- Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, 23(4), 441–474.
- Sereno, S., O'Donnell, P., & Rayner, K. (2006). Eye movements and lexical ambiguity resolution: Investigating the subordinate bias effect. *Journal of Experimental Psychology: Human Perception and Performance*, 32, 335–350.
- Troy, M., Chen, S. C., & Stern, J. A. (1972). Computer analysis of eye movement patterns during visual search. *Aerospace Medicine*, 42(4), 390–394.
- Weir, C.J. (2005). *Language testing and validation*. New York: Palgrave Macmillan.
- Williams, R. S., & Morris, R. K. (2004). Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology*, 16, 312–339.
- Yamashita, J. (2003). Processes of taking a gap-filling test: Comparison of skilled and less skilled EFL readers. *Language Testing*, 20(3), 267–293.
- Yi'an, W. (1998). What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple-choice task. *Language testing*, 15(1), 21–44.

British Council Assessment Research Awards and Grants

If you're involved or work in research into assessment, then the British Council Assessment Research Awards and Grants might interest you.

These awards recognise achievement and innovation within the field of language assessment and form part of the British Council's extensive support of research activities across the world.

LOOKING INTO TEST-TAKERS' COGNITIVE PROCESSES WHILE COMPLETING READING TASKS:

A mixed-method eye-tracking and stimulated recall study

AR-G/2015/001

**Tineke Brunfaut and
Gareth McCray**
Lancaster University, UK

**ARAGs RESEARCH REPORTS
ONLINE**

ISSN 2057-5203

© British Council 2015

The British Council is the
United Kingdom's international
organisation for cultural relations
and educational opportunities.

