

**LOOKING INTO READING II: A FOLLOW-UP STUDY  
ON TEST-TAKERS' COGNITIVE PROCESSES WHILE  
COMPLETING APTIS B1 READING TASKS**

VS/2016/001

**Tineke Brunfaut, Lancaster University**

## ABSTRACT

---

This study investigated 25 ESL test-takers' cognitive processing while completing a set of *Opinion Matching* tasks designed and piloted for use on the Aptis reading test and targeting the CEFR-B1 level. Insights were gained through the recording of participants' eye traces during task completion, immediately followed by a stimulated recall after each task in which participants described, in their first language, how they had completed it.

The study follows up on the findings of Brunfaut & McCray (2015), who investigated test-takers' task processing on the full Aptis reading test, and found support for the construct validity of the test. However, a somewhat weaker alignment between the intended and actual reading processes used by test-takers was found for the B1 banked gap-fill tasks of the Aptis reading test. The aim of this follow-up study, therefore, was to explore the cognitive processing on an alternative, newly designed set of *Opinion Matching* tasks to be able to evaluate the extent to which the new tasks elicited the specified processes for this level of the Aptis reading test (see O'Sullivan & Dunlea, 2015).

Test-takers were found to use a wide range of cognitive processes while they were completing the B1 *Opinion Matching* tasks, covering both lower- and higher-level processes as defined in Khalifa & Weir's (2009) model of reading. Most often, when they successfully solved the item, the test-takers had adopted a careful and global reading approach (and sometimes they also did some expeditious reading), and they had combined lexical processing and/or meaning-making within sentences, across sentences or at textual and intertextual levels with inferencing. This global comprehension approach and the extensive engagement with higher-level processing associated with the B1 *Opinion Matching* items not only differs substantially from the local and lower-level processing patterns associated with the original B1 banked gap-fill items, but also suggests a suitable match with the intended processes specified by the test developers for the Aptis B1 reading tasks.

## Author

---

**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*, *Language Assessment Quarterly* and *Language Testing*. Tineke has also been involved in test development in a range of languages and countries around the world.

## Acknowledgements

---

I would like to express my gratitude to the British Council for funding this research. A big thank you also goes to my research assistants, Diana Mazgutova, Gareth McCray, Anchana Rukthong, and Pucheng Wang, and to the participants.

# CONTENTS

---

<b>1. INTRODUCTION</b>	<b>5</b>
<b>2. THEORETICAL BACKGROUND</b>	<b>6</b>
<b>3. RESEARCH QUESTIONS</b>	<b>8</b>
<b>4. METHODOLOGY</b>	<b>8</b>
4.1 Participants	8
4.2 Materials	9
4.2.1 Reading tasks	9
4.2.2 Full Aptis test	10
4.3 Data collection methodology and procedures	10
4.4 Ethical procedures and consent	12
4.5 Data analyses	12
4.5.1 Eye-tracking analyses	13
4.5.2 Stimulated recall analyses	14
<b>5. FINDINGS</b>	<b>16</b>
5.1 Descriptive statistics	16
5.2 Eye-tracking	17
5.2.1 Eye-tracking findings on cognitive processes when completing Aptis B1 Opinion Matching reading tasks (RQ1) and comparisons with the original Aptis B1 Banked Gap-fill reading tasks (RQ1a)	17
5.2.2 Eye-tracking findings on cognitive processes depending on test-takers' L2 (reading) proficiency (RQ1b & RQ1c)	23
5.3 Stimulated recall	25
5.3.1 Stimulated recall findings on cognitive processes during Aptis B1 Opinion Matching reading task completion (RQ1)	25
5.3.2 Stimulated recall findings on differences in cognitive processes between the 'old' and 'new' B1 task type	29
5.3.3 Stimulated recall findings on cognitive processes depending on test-takers' L2 (reading) proficiency (RQ1b & RQ1c)	30
<b>6. DISCUSSION</b>	<b>33</b>
<b>7. CONCLUSION</b>	<b>34</b>
<b>REFERENCES</b>	<b>35</b>

## LIST OF TABLES

Table 1: Eye-tracking metrics (Brunfaut & McCray, 2015, p.18)	13
Table 2: Adapted coding framework stimulated recalls	15
Table 3: Descriptive statistics – Aptis B1 Opinion Matching reading tasks used for eye-tracking and stimulated recall (n=25)	16
Table 4: Descriptive statistics – full Aptis system (n=25)	16
Table 5: Aptis components – CEFR levels of participants (n=25)	16
Table 6: Descriptive statistics – Eye-tracking measures for the Aptis B1 Opinion Matching tasks	19
Table 7: Results Aptis B1 Opinion Matching eye-tracking analyses compared to Brunfaut & McCray (2015) study's results (RQ1a)	20
Table 8: Eye-tracking support for RQ1a hypotheses	21
Table 9: Results eye-tracker analyses in relation to L2 reading proficiency (RQ1b) and overall L2 proficiency (RQ1c)	23
Table 10: Eye-tracking support for RQ1b and RQ1c hypotheses	24
Table 11: Stimulated recall results on cognitive processes during Aptis B1 Opinion Matching reading task completion (RQ1)	26
Table 12: Stimulated recall results on cognitive processes when correctly completing B1 banked gap-fill versus B1 Opinion Matching items (RQ1a)	29
Table 13: Stimulated recall results on cognitive processes of correct items depending on test-takers' L2 reading proficiency (RQ1b)	31
Table 14: Stimulated recall results on cognitive processes of correct items depending on test-takers' L2 proficiency (RQ1c)	33

## LIST OF FIGURES

Figure 1: Khalifa & Weir's model of cognitive processing in reading – adapted from Khalifa & Weir (2009, p. 43) (Brunfaut & McCray, 2015, p. 7)	6
Figure 2: Key eye-tracking measures (Brunfaut & McCray, 2015, p. 10)	7
Figure 3: Adapted Opinion Matching task layout	9
Figure 4: Flowchart of the first data collection session	11
Figure 5: Heat maps of Aptis B1 Opinion Matching reading tasks	18
Figure 6: Heat maps of the old Aptis B1 Banked Gap-fill tasks (Brunfaut & McCray, 2015, p. 28)	18
Figure 7: Reported combined uses of processes during B1 Opinion Matching completion	28

# 1. INTRODUCTION

---

This study examined the cognitive processing of 25 test-takers while completing Aptis B1 reading tasks. The study follows up on the findings of the 'Looking into Reading' report by Brunfaut & McCray (2015), which investigated test-takers' task processing on the full Aptis reading test by means of eye-tracking and stimulated recall methodologies. The Brunfaut & McCray (2015) study concluded 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 & Weir's (2009) model of reading, providing evidence of construct validity of the test for the sample of test-takers. Different patterns were observed in the main forms of processing used to complete the four different CEFR-linked Aptis reading task types (A1, multiple-choice gap-fill; A2, sentence ordering; B1, banked gap-fill; and B2, matching headings) and these differences appeared to be closely related to the task type. Generally, these patterns roughly matched the intended target processes for each CEFR-linked task level, as defined in the Aptis General Technical Manual, Version 1.0 (O'Sullivan & Dunlea, 2015). An exception to this, however, was test-takers' processing of the Aptis B1 banked gap-fill reading items. Although these B1 tasks were found to elicit a range of reading comprehension processes, the reading processes test-takers used did not necessarily align with the expected processes (as outlined in O'Sullivan & Dunlea, 2015).

In response to the findings of Brunfaut & McCray (2015), the Aptis test developers explored alternative task types to ensure assessment of the intended cognitive processes for the B1-target reading task. In practice, a matching task type – the *Opinion Matching* task – was chosen, and a number of versions of the task were developed and piloted on the target population to screen for general functioning and quality of the task type.<sup>1</sup> Subsequently, the British Council commissioned a study to investigate the cognitive processing of test-takers while completing the new B1 reading tasks, in order to evaluate the extent to which the new tasks elicited the specified processes for this level of the reading test (see O'Sullivan & Dunlea, 2015). The present report describes this follow-up study and its findings.

The follow-up study, entitled 'Looking into Reading II', aimed to examine test-takers' cognitive processing while completing the new Aptis B1 reading task type, i.e. the *Opinion Matching* task. The study investigated test-takers' task processing both in general and also relative to their L2 proficiency (as measured by the full Aptis test) and L2 reading proficiency (as measured by the Aptis reading component). In order to investigate this, and in a similar vein to the first 'Looking into Reading' study reported in Brunfaut & McCray (2015), a combination of eye-tracking and retrospective verbal protocols, with eye-tracking traces as recall enhancing stimuli, were used. The data, in combination with information on the processes intended to be assessed, provide key information on the validity of the *Opinion Matching* task.

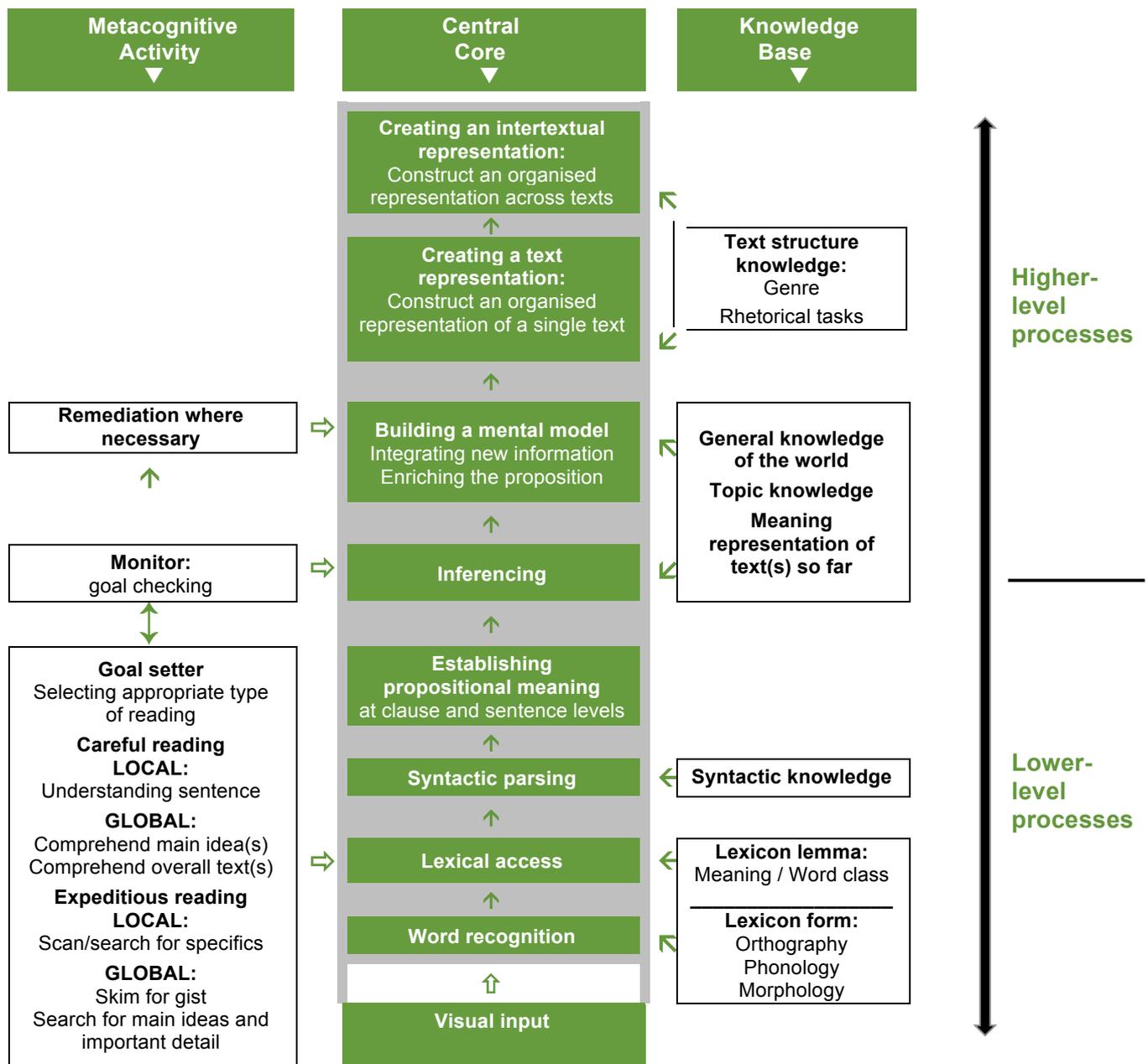
---

<sup>1</sup> In what follows, the Aptis B1 banked gap-fill task investigated in Brunfaut & McCray (2015) is occasionally referred to as the 'old/original B1 task', whereas the Aptis B1 *Opinion Matching* task analysed in this study is referred to as the 'new B1 task'.

## 2. THEORETICAL BACKGROUND

In line with the first 'Looking into Reading' study (Brunfaut & McCray, 2015), Khalifa & Weir's (2009) model of reading formed the theoretical foundation of this follow-up study due to the fact that it aligns with the socio-cognitive approach to test development and validation which has been adopted by the Aptis test developers (O'Sullivan, 2015). Thus, the inferences from the empirical data in this study on test-takers' reading processes during *Opinion Matching* completion have been drawn with reference to the reading processes as theorised by Khalifa & Weir (2009). For ease of reference, the model is reproduced here in Figure 1, but for definitions and more detailed information on its components and the various processes, refer to Brunfaut & McCray (2015, pp. 5–6).

Figure 1: Khalifa & Weir's model of cognitive processing in reading – adapted from Khalifa & Weir (2009, p. 43) (Brunfaut & McCray, 2015, p. 7)



Empirical insights into reading and test completion increasingly come from mixed-methods research. In particular, the need to explore test-takers' reading processes in a direct manner – not just relying on expert judgments of what readers might do – has been emphasised in the literature (e.g., Alderson, 2000). Therefore, in addition to quantitative analyses of reading test performance results, valuable insights into the nature of reading and the testing of reading have been gained through the use of qualitative methods such as verbal reports (e.g., Israel, 2015; Pressley & Afflerbach, 1995). Recently, language testing researchers have started experimenting with the use of eye-tracking technology to investigate test constructs and test validity. The combined use of eye-tracking and stimulated recalls, whereby test-takers are prompted to recall their thought processes by means of watching their own eye-movements as they responded to an item, has shown to be particularly useful (see e.g., Bax, 2013; Brunfaut & McCray, 2015). Therefore, as in the first study (Brunfaut & McCray, 2015), this study set out to investigate test-takers' cognitive processes during reading test completion by means of eye-tracking and stimulated recall methodologies. For a description and review of eye-tracking methodology and a rationale for its use for the present research purpose, refer to Brunfaut & McCray (2015, pp. 8–10). To aid the reading of this current research report, however, this section very briefly explains four key terms related to eye movements and eye-tracking methodology: saccades, fixations, regressions, and areas of interest.

When we read in English, our eyes do not move continuously across the lines, but they make series of forward *saccades* – small forward jumps from left to right going along the lines. The length of the forward saccades our eyes make may vary depending on issues such as the cognitive load involved in the task, the type of reading, or the proficiency of the reader. Our eyes may make several saccades while focused on one word, but they may also 'bridge' a number of words in one go. Sometimes, our eyes also make backward saccades from right to left, termed *regressions*. Regressions may happen for a number of reasons. Two common reasons for their occurrence while reading are, firstly, because our eyes 'overshot' an area we wanted to visually process and thus our brain 'self-corrects' to bring the area back into focus or, secondly, because of a comprehension breakdown we are attempting to repair, we need to move backwards and re-read a section of text. The number of regressions our eyes make may be associated with factors such as text difficulty and reading ability. At the end of each forward or backward movement, our eyes make a *fixation* – a short pause during which our eyes rest at a specific point on the page/screen in which information is taken in for processing. The time taken to fixate – the fixation duration – may also vary depending on factors such as word familiarity, lexical ambiguity, plausibility, the type of reading, etc. A visual representation of these eye movement characteristics is provided in Figure 2.

Figure 2: Key eye-tracking measures (Brunfaut & McCray, 2015, p. 10)



When working within a reading testing context, as opposed to reading *per se*, there are different aspects to a task that a test-taker is likely to process during test completion, for example, the task instructions, the reading input texts, and the items or questions asked to assess the test-taker's comprehension of the textual input. Differences in test-taker ability, for example, may be associated with differences in time spent on different parts of the visual stimulus or switches between the parts. These different parts can be defined by the researcher on the basis of the study's focus, and are called *areas of interest* (AOIs).

The use of eye-tracking methodology to explore cognitive processing in reading is based on Rayner's (1998) position that there is a close link between the text our brain is processing and the point on the page (or screen) on which our eyes are focusing. Evidence for the types of eye movements described above and their characteristics has been found in empirical research on first language reading and a growing body of studies on second language reading (see Brunfaut & McCray (2015) for a review).

## 3. RESEARCH QUESTIONS

---

The aims of this study, as stated in section 1, were formulated as the following research questions.

**Overarching question:**

**RQ1.** What cognitive processes do test-takers employ during completion of the Aptis B1 *Opinion Matching* reading task?

**Sub-questions:**

**RQ1a.** Are there any differences in cognitive processes depending on the task type (i.e. the new B1 *Opinion Matching* task as compared to the original Aptis B1 banked gap-fill task)?

**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 Aptis test components?

## 4. METHODOLOGY

---

Given the effective methodological design of the first Aptis reading study (Brunfaut & McCray, 2015), this follow-up investigation similarly combined eye-tracking and stimulated recall methodology to obtain data on test-takers' cognitive processes during completion of the Aptis B1 *Opinion Matching* reading task. This also enabled comparisons between the two studies (i.e., Looking into Reading I & II – Brunfaut & McCray (2015) and the present study), and specifically between the 'new' and the 'old' Aptis B1 reading tasks. Also, to further ensure comparability of the two studies, this follow-up study recruited participants similar in profile to the first study's participant group, followed similar data collection procedures, and adopted the same data analytical approaches.

### 4.1 Participants

The participants in this study were 25 English as a Second Language (ESL) speakers from three different first language backgrounds: 10 Thai-L1, 10 Chinese-L1, and 5 Russian-L1 speakers.

In terms of gender, 44% were male and 56% were female. Their ages ranged between 18 and 29 years old (M=21.5).

Four percent of the participants were enrolled on a pre-session English language course, 68% were undergraduate and 28% were postgraduate students at a British university. They had been living in English-speaking countries for between half a year and eight years (M=2.0 years).

## 4.2 Materials

### 4.2.1 Reading tasks

To investigate the cognitive processes when completing the new Aptis B1 task, the *Opinion Matching* task, the British Council provided three versions of the new task type which had previously been piloted. One task was used as a sample task for participant familiarisation and two tasks were used for data collection on participants' cognitive processing. This matched the size of the dataset of the first study: the Brunfaut & McCray (2015) study similarly looked into processing on two versions of each CEFR-target level task (so two B1 banked gap-fill tasks). Each *Opinion Matching* task consisted of seven items, thus data was collected on a total of 14 items per participant.

The *Opinion Matching* task consists of four (4) short texts, each one paragraph in length, which reflect four people's interview replies on a common theme. The shared topic and a description of the publication outlet are provided to the test-takers in the task instructions. The task requires the test-takers to match each of a set of seven (7) questions to one of the interviewees, on the basis of the information provided in the texts.

Given the use of eye-tracking and stimulated recall methodology in the study, the Aptis task presentation was mirrored in an offline format to ensure compatibility with the eye-tracking software, and to allow for pausing and replays of eye traces for the purposes of the stimulated recall. To maintain ecological validity, only minimal formatting changes were made to the task presentation to enable the interpretation of eye traces. On the British Council's test delivery platform, the Aptis *Opinion Matching* presentation allows for scrolling up and down the text, however, in this study the stimulus was presented on a single screen. This is because moving screens interfere with interpreting the link between a particular eye trace and what the participant was focusing on at that point. Therefore, the questions and all text paragraphs were presented together so that no scrolling up and down was needed. The screenshot in Figure 3 shows the task layout as designed for the purposes of the study.

Figure 3: Adapted *Opinion Matching* task layout



Note: This figure shows the layout of the task. For reasons of confidentiality, the text has been obscured.

## 4.2.2 Full Aptis test

To obtain a measure of participants' overall English language proficiency and their English reading proficiency, the full computer-based Aptis test – consisting of the components grammar/vocab, reading, listening, writing, and speaking – was administered. This was done to enable the exploration of potential differences in cognitive processing depending on test-takers' L2 reading proficiency and overall L2 proficiency (RQs1b & 1c).

The same version of the test used in the first study (Brunfaut & McCray, 2015) was used again to allow for comparison across projects and to avoid duplication of items between the full Aptis and the specific reading tasks used for eye-tracking. The regular Aptis test administration procedures stipulated by the British Council were strictly adhered to.

## 4.3 Data collection methodology and procedures

The data collection protocols were kept as similar as possible to those in the first study (Brunfaut & McCray, 2015), i.e. the data were gathered in two phases.

### Phase 1

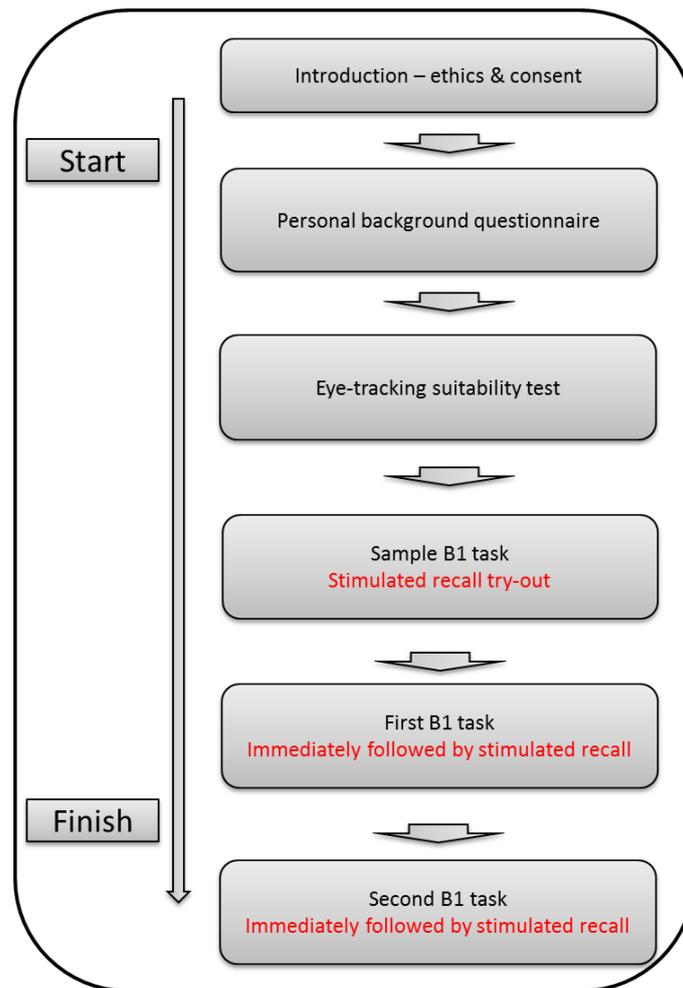
The aim of the first phase was to collect data that would inform the response to the overarching research question (RQ1 – *What cognitive processes do test-takers employ during completion of the Aptis B1 Opinion Matching reading task?*) and the first sub-question (RQ1a – *Are there any differences in cognitive processes depending on the task type, i.e. the B1 Opinion Matching task as compared to the original Aptis B1 banked gap-fill task?*)

In this phase, the participants completed the *Opinion Matching* tasks while their eye traces were being recorded. The eye traces were collected using a Tobii TX300 eye-tracker, an unobtrusive, high-precision eye-tracker (300Hz sampling rate, accuracy 0.4°). The tasks were displayed in the Verdana font with a font size of 24px/18pt on a 23" monitor with an aspect ratio of 16:9 and a resolution of 1920x1080.

Each task was immediately followed by a retrospective verbal report of participants' cognitive processes during task completion, i.e. the participants were prompted to verbalise the thoughts they had during task completion. To assist them with recall, they were shown screen recordings of their eye traces from when they had been completing the tasks (i.e. the eye traces functioned as stimuli for retrospection).

The procedures are outlined in Figure 4. Due to the methods used, the data were collected from one participant at a time. The session lasted approximately 30–40 minutes per participant.

Figure 4: Flowchart of the first data collection session



As visualised in Figure 4, each participant was first given an introduction to the nature and purposes of the study. Then, the participant was asked for their written consent and asked to fill out a paper-based participant background questionnaire. This was followed by a technical eye-tracking suitability test to determine whether the participant's eye-traces could sufficiently be captured by the hardware.<sup>2</sup> In practice, all participants met the conditions for inclusion in the study. After this, the eye-tracking and stimulated recall procedure began.

First, the participant was asked to complete an example *Opinion Matching* task to familiarise themselves with the task type. At the same time, the participant's eye traces were recorded. When the participant had completed the sample task, the stimulated recall procedure was trialled. The screen recordings with eye traces overlaid were replayed to the participant and the participant was asked to recall and verbalise his/her task completion processes. If necessary, feedback was given to the participant on the functioning of the task to aid their understanding of what was expected, or on how to conduct the recall. Five participants had assumed that some of the questions served as distractors and thus only answered four questions at first, whereas all seven need to be completed. The latter was pointed out to these participants by a research assistant.

<sup>2</sup> Eye-tracking suitability tests are run to prevent issues such as long eyelashes, droopy eyelids, or verifocal glasses interfering with the recording of eye traces in a manner unsuitable for data analyses (Holmqvist et al., 2011).

After the sample task, the main data were collected. The participant was asked to complete two *Opinion Matching* tasks while their eye traces were simultaneously recorded. Each completed task was followed by an immediate replay of the eye traces, pausing after each item completion attempt or when a participant felt suitable, and a request to verbalise was given as to how they approached the reading task and items in general, what they had been thinking during task completion, and how they had arrived at each of the answers they gave. For consistency, all stimulated recalls were conducted following a script with instructions and questions posed by the researcher. The stimulated recalls were audio- and video-recorded to enable and facilitate the understanding and interpretation of the stimulated recall data at the analysis stage.<sup>3</sup>

To ensure that the participants would be able 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. The recall data were collected by three research assistants who had the same L1 as the participants (Thai, Chinese or Russian). These assistants were all linguists, specialised in language testing and second language acquisition (SLA), and had also conducted the recalls in the Brunfaut & McCray (2015) study. They were given refresher training for the purposes of this follow-up study. Furthermore, to generate high-quality eye-tracking data, calibration was refreshed before each individual task.

## Phase 2

During the second data collection session, the same participants that took part in the eye-tracking/stimulated recall session were administered the full Aptis test (all five components). This was 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 relative to 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 researcher's institution. The session was supervised by the researcher, and the official Aptis test's procedures were strictly adhered to.

Prior to the main data collection, all instruments and procedures were piloted with two people.

## 4.4 Ethical procedures and consent

Adhering to the regulations at the researcher's institution, ethical approval for the study was sought and granted. The participants were provided with a written information sheet. In addition, the nature of the study, the participant's involvement, and the contact details of the researcher and Head of Department were also explained orally. All participants gave their consent in writing.

## 4.5 Data analyses

Based on the findings and experiences of the first study (Brunfaut & McCray, 2015), data from eye-tracking and stimulated recall can be triangulated and can also be combined to address weaknesses inherent in either method. For example, eye-tracking data is useful to shed light on low-level processes such as word recognition speed, whereas stimulated recalls can provide information on test-takers' use of higher level processes such as inferencing. In what follows, brief descriptions of the data analysis approaches for both data sources – eye movements and stimulated recalls – are given.

---

<sup>3</sup> Based on previous research experiences of stimulated recalls with visual stimuli, participants tend to physically point at the stimulus when recalling their processing. Video-capturing the research process can, therefore, facilitate the understanding and interpretation of the stimulated recall data at the analysis stage.

## 4.5.1 Eye-tracking analyses

### Measures

For reasons of comparability between the 'new' and the 'old' B1-target task, the same eye-tracking measures were looked into and the same analyses were conducted as in Brunfaut & McCray (2015). These can be summarised as follows<sup>4</sup>: the data were analysed according to 11 eye-tracking metrics which relate to the test-takers' fixations, saccades and regressions, and which can be subdivided into three processing-type groups – global processing, text processing, and task processing.<sup>5</sup> Fixations were determined using the Tobii I-VT velocity and acceleration-based filter with its default settings. Table 1 provides an overview of the measures and their technical definitions.

Table 1: Eye-tracking metrics (Brunfaut & McCray, 2015, p.18)

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 left 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 right 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 Aoi switches between text and response options	The number of movements between Areas of Interest (Aois) containing text and an Aoi containing the response options.

\*Scaled for font size (see below).

<sup>4</sup> For more detailed info on the analyses and their rationale, see Brunfaut & McCray (2015).

<sup>5</sup> Global processing measures are calculated on the basis of eye-movement data on both the text and items (questions). Text processing measures relate to the input texts only (the four texts, but not the questions), whereas task processing measures concern the interactions between texts and items, or the items only (the seven questions).

### **Font scaling**

To fit the texts and items of each task onto a single slide for presentation on the eye-tracker, the font sizes had to be set to a slightly different standard to that of some of the task types (A1, A2, B2) used in Brunfaut & McCray (2015). To allow for comparisons across all Aptis CEFR-linked reading task types, some measures were scaled. Namely, distance measures which would reduce with a smaller font size (indicated with \* in Table 1), were scaled by 1.07 for the new B1 *Opinion Matching* tasks (which is the same as the old B1 banked gap-fill tasks).

### **Analyses**

To explore test-takers' processing while completing the Aptis B1 *Opinion Matching* reading tasks (RQ1) and potential differences in processing to the original Aptis B1 banked gap-fill reading tasks (RQ1a), the eye-tracking data collected on the two *Opinion Matching* tasks were visually represented in heat maps and analysed according to the 11 metrics listed in Table 1. Differences in eye movements with the old B1 banked gap-fill tasks, as well as with the other Aptis reading tasks (A1, A2, B2) were established through Mann-Whitney U tests.

In order to investigate differences in test-takers' cognitive processing depending on their English L2 reading proficiency (RQ1b) and their overall English L2 proficiency (RQ1c), Spearman's correlations were run between each of the 11 eye-tracking measures (from test-takers' eye movements during *Opinion Matching* task completion) and the measures of test-takers' L2 (reading) proficiency.

## 4.5.2 Stimulated recall analyses

To analyse participants' verbal reports of their task completion processes, the same methodology was followed as that in Brunfaut & McCray (2015).<sup>6</sup> The verbal protocols, which had been video- and audio-recorded, were transcribed and translated from the participants' L1 into English by the research assistants. The English transcripts were then coded by the principal investigator, using the qualitative data analysis software Atlas.ti v7, adopting Khalifa & Weir's (2009) model of cognitive processing in reading as the basis of the coding framework. In addition, three extra codes were added to the coding framework used in Brunfaut & McCray (2015, p. 22), on the basis of observations made on the nature of the data during the coding process. Specifically, the codes 'Reading instruction info' and 'Pre-reading item questions' were added since a number of participants explicitly mentioned starting by reading the information given in the instruction and/or by reading the questions of the *Opinion Matching* task to gain an overall idea of the texts' topic, text type, content and/or what to pay attention to while reading. The code 'Test-taking strategy' was developed to capture some participants' reasoning regarding the number of times an interviewee's text can/should be matched to the seven questions (e.g. at least once). The code 'Creating paragraph level representation', which had been specifically created to capture processing on the multi-paragraph texts of the B2 matching headings task in the Brunfaut & McCray (2015) study was dropped from the coding framework because the B1 *Opinion Matching* task consists of four separate text excerpts (texts in their own right) each only one paragraph in length, thus rendering this code irrelevant. The adapted coding framework is provided in Table 2.

The coding framework was applied to items answered correctly, as well as items answered incorrectly, with an additional tag of 'W' for wrongly answered items. This distinction was made for the purpose of test validation analyses, whereby it is vital to know what leads to a correct answer or what may cause construct-irrelevant variance.

---

<sup>6</sup> For more detailed info on the analyses and their rationale, see Brunfaut & McCray (2015).

Table 2: Adapted coding framework simulated recalls

Goal setting codes	Central core codes	Additional codes
Careful reading – local	Word recognition	
Careful reading – global	Lexical access	
Expeditious reading – skimming	Syntactic parsing	Collocation
Expeditious reading – search reading	Establishing propositional meaning	
Expeditious reading – scanning	Inferencing	Background knowledge
	Building a mental model	
	Creating text level representation	
	Creating intertextual representation	
		Pure guess Test-taking strategy Reading instruction info Pre-reading item questions

Once the data were coded, the frequency of codes was calculated to establish what cognitive processes test-takers employ while completing B1 *Opinion Matching* tasks (RQ1). Furthermore, a number of subanalyses were conducted to inform the answers to RQs1a-1b-1c. Namely, to evaluate the effectiveness and validity of the B1 *Opinion Matching* task, the codings were tallied and compared with those of the 'old' B1 banked gap-fill task (RQ1b).

Additionally, to investigate processing differences depending on test-takers' L2 (reading) proficiency (RQ1b & RQ1c), the stimulated recall data were split into groups according to the participants' L2 (reading) ability, expressed in CEFR proficiency levels and measured by means of the full Aptis test. The codings were then analysed for each of the (reading) proficiency groups, and also compared across groups. The subanalyses involved the processes associated with correctly-answered items, since – from a validation perspective – these should reflect the intended construct. However, the processes associated with incorrectly-answered items were also examined to provide a richer interpretation of test-takers' cognitive processing during task completion.

## 5. FINDINGS

### 5.1 Descriptive statistics

Participants' performance results on the Aptis B1 *Opinion Matching* readings tasks are presented in Table 3. The mean score and scoring range indicate that the participants performed well on these tasks (on which participants' eye movements were recorded and in relation to which they produced the stimulated recalls).

Table 3: Descriptive statistics – Aptis B1 *Opinion Matching* reading tasks used for eye-tracking and stimulated recall (n=25)

	Max. possible score	Min.	Max.	M	SD
<b>B1 Opinion Matching tasks</b>	14	8	14	12.36	1.66

Table 4 provides descriptive statistics for participants' performances on the full Aptis test which was administered to obtain an L2 proficiency and L2 reading proficiency measure for the participants.

In Table 5, these results have been translated into the number of participants performing at a particular CEFR level, as stated in the score reports retrieved from the British Council. The table shows that, similar to the first study (Brunfaut & McCray, 2015), the volunteer group willing to participate can be characterised as mostly "independent and proficient users" (Council of Europe, 2001).

Table 4: Descriptive statistics – full Aptis system (n=25)

	Max. possible score	Min.	Max.	M	SD
<b>All skill components</b>	200	148	192	170.08	12.79
<b>Grammar &amp; vocab</b>	50	34	46	39.12	3.31
<b>Reading</b>	50	24	50	44.88	5.86
<b>Listening</b>	50	26	48	43.68	4.50
<b>Speaking</b>	50	26	48	37.68	5.78
<b>Writing</b>	50	32	50	43.84	5.74

Table 5: Aptis components – CEFR levels of participants (n=25)

Aptis Component	A1	A2	B1	B2	C
<b>Reading</b>	0	1	0	4	20
<b>Listening</b>	0	0	1	0	24
<b>Speaking</b>	0	1	7	15	2
<b>Writing</b>	0	0	4	7	14

To gauge the comparability of the first study (Brunfaut & McCray, 2015) and the present study, an independent-samples t-test was conducted to compare the L2 proficiency of the participant groups. No significant differences were found in full Aptis scores of the first study's participants ( $M=166.72$ ,  $SD=12.89$ ) and the present study's [ $M=170.08$ ,  $SD=12.79$ ,  $t(48)=-.93$ ,  $p=.36$ ], or in Aptis reading scores of the first study's participants ( $M=42.32$ ,  $SD=5.09$ ) and the present study's [ $M=44.88$ ,  $SD=5.86$ ,  $t(48)=-1.65$ ,  $p=.11$ ]. The magnitude of the differences in the means was small ( $\eta^2_{\text{full Aptis}}=.018$ ;  $\eta^2_{\text{Aptis reading}}=.054$ ) (following Cohen, 1988). In addition to the fact that the participants in the present study were strictly recruited according to the same test-taker characteristics profile of the first study, these statistical results provide support for the validity of making comparisons between the two studies (in particular RQ1a).

## 5.2. Eye-tracking

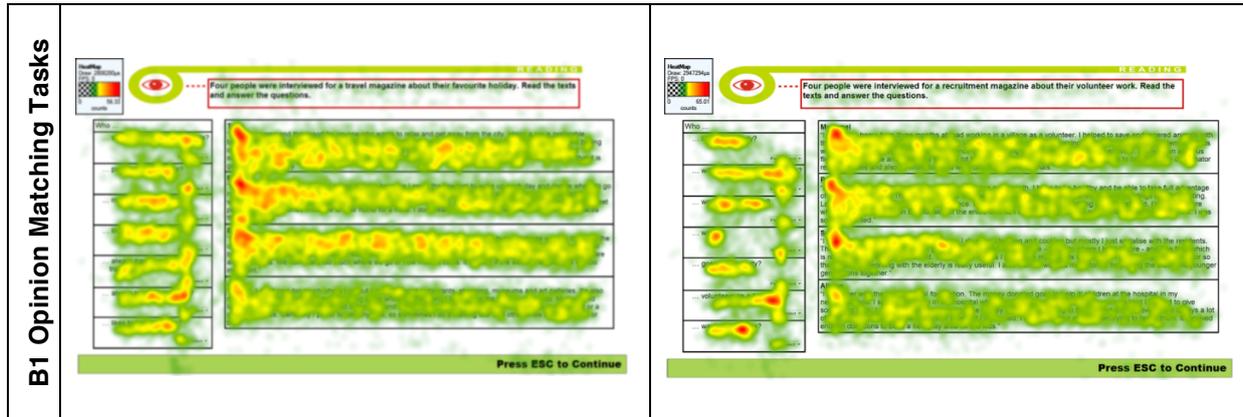
Two types of data were collected to help uncover test-takers' cognitive processing while completing Aptis B1 *Opinion Matching* reading tasks: eye-movement data and verbal reports of test-takers' thought processes. This section presents the results of the eye-movement analyses.

### 5.2.1 Eye-tracking findings on cognitive processes when completing Aptis B1 Opinion Matching reading tasks (RQ1) and comparisons with the original Aptis B1 Banked Gap-fill reading tasks (RQ1a)

Initial insights into test-takers' cognitive processing while completing Aptis B1 *Opinion Matching* reading tasks (RQ1) were gained from heat map inspection. Heat maps are visualisations resulting from participants' eye movements and which plot the aggregate amount of time participants spent focusing on particular areas of the input. The colours – ranging from transparent (no fixations) over green and through yellow to red – indicate increasing amounts of time visually focusing on an area. Figure 5 shows these visualisations for all items of the two *Opinion Matching* tasks, with each participant's data carrying the same weight.

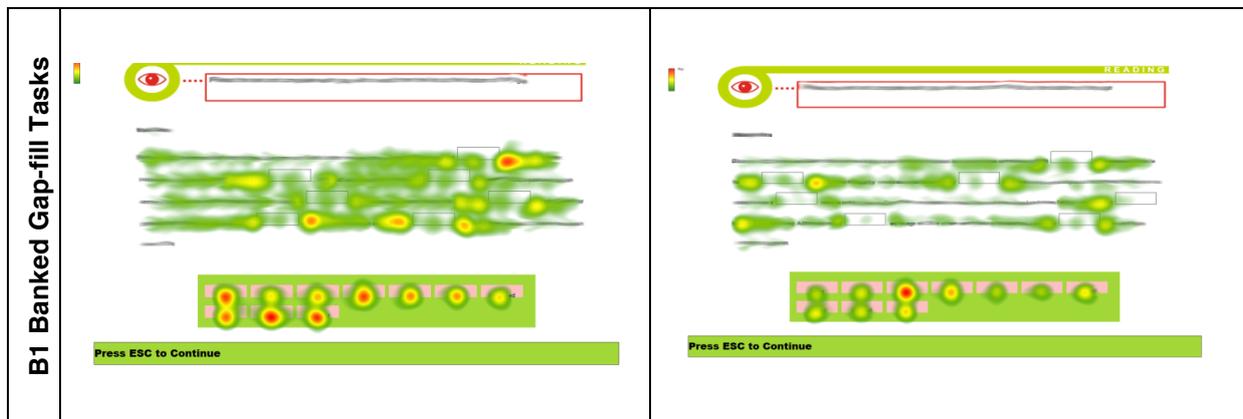
The pattern that seems to emerge from these heat maps is that participants' attention covered the full four texts as well as the questions. Overall, attention within the texts seems to be quite evenly spread or at least not focused on isolatable words/phrases, with the exception of more attention to the area where the names of the people associated with each text were printed (which is needed to answer the questions). This suggests a careful and more global reading approach. Also, the initial three texts in each task appear to have received more attention overall as compared to the last (fourth) text. Potentially, this is because participants started by reading the texts in order of presentation while at the same time solving questions as they went along. Thus, they had to consider seven questions in relation to the first text, were then solving questions along the way, and thus most likely had to consider fewer information points by the time they reached the fourth text.

Figure 5: Heat maps of Aptis B1 Opinion Matching reading tasks



For comparison (RQ1a), the heat maps of the old B1 banked gap-fill tasks showed a rather different picture (see Figure 6): test-takers' attention seemed to particularly be drawn to the words surrounding the gaps, which suggested more careful local reading and lower-level processing (Brunfaut & McCray, 2015).

Figure 6: Heat maps of the old Aptis B1 Banked Gap-fill tasks (Brunfaut & McCray, 2015, p. 28)



More detailed information on participants' cognitive processing was obtained through in-depth sub-analyses of the eye trace data in terms of the 11 measures listed in Table 1. Descriptive statistics for each of the measures are provided in Table 6 (IQR=Interquartile range).

Table 6: Descriptive statistics – Eye-tracking measures for the Aptis B1 Opinion Matching tasks

		Global processing measures		Text processing measures							Item processing measures	
		Total number of fixations	Total fixation time on text and responses (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
B1 Opinion Matching	Median	749	222	413	107	193	-70	0.25	210	0.43	47	0.20
	IQR	306	98	173	30	67	27	0.07	23	0.18	25	0.04
	Min	495	119	256	67	84	-110	0.14	167	0.23	24	0.10
	Max	1321	368	759	179	370	-47	0.37	240	0.74	89	0.24

Table 7 presents the results of the eye-movement analyses of the Aptis B1 *Opinion Matching* tasks, i.e. the “New B1 items”, together with the results of the eye-movement analyses of all Aptis reading tasks investigated in the first study (Brunfaut & McCray, 2015) – A1 multiple-choice gap-fill, A2 sentence ordering, B1 banked gap-fill, and B2 matching tasks. As was the case in the first study, differences can be observed in the eye-tracking measures between the B1 *Opinion Matching* tasks and the A1-, A2- and B2-target level tasks. In addition, the results for the new B1 *Opinion Matching* task (in green print) are noticeably different from those of the first study’s B1 banked gap-fill tasks (in red print). Significance testing of these differences was conducted via the Mann-Whitney U tests, which confirmed that the new B1 tasks resulted in different eye-movement patterns from the old B1 tasks in almost all respects (in orange print) (RQ1a).

Table 7 shows that, with the exception of the measures *Median length of regressions* and *Median fixation duration*, there are statistically significant differences between the two B1-target level task types for all the measures:

- *global processing measures*: ‘total number of fixations’ and ‘total fixation time spent on text and responses’
- *text processing measures*: ‘number of forward saccades’, ‘median length of forward saccades’, ‘number of regressions’, ‘proportion of regressive movements’, ‘sum of fixation time on text per word’
- *item processing measures*: ‘number of Aol switches between text and responses’ and ‘proportion of time spent fixating on responses’.

Whereas in the first study, it was found that when participants were processing the texts, the old B1 tasks elicited several eye-movement patterns similar to those when processing the A1 tasks – indicating more local, careful reading around the gaps – the present study’s new B1 tasks show mostly significantly different results on the eye-tracking metrics as compared to the old B1 tasks. This suggests different cognitive processing of the B1 *Opinion Matching* tasks than the B1 banked gap-fill tasks, specifically a tendency towards less local and more global careful reading.

Table 7: Results Aptis B1 Opinion Matching eye-tracking analyses compared to Brunfaut & McCray (2015) study's results (RQ1a)

		Global processing measures		Text processing measures							Item processing measures	
		Total number of fixations	Total fixation time on text and responses (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
First study	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
	Old 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
This study	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
	New B1 Items	749	222	413	107	193	-70	0.25	210	0.43	47	0.20
	B2 Items	859	280	606	77.95	189	-64.77	0.19	237	0.29	72	0.28
Significance tests	A1 – new B1	***	***	***	***	***		***		***	***	***
	A2 – new B1	***	***	***	***	***	*			***		*
	Old B1 – New B1	***	***	***	***	***		***		***	***	***
	B2 – new B1		***		***		*	*	***	***	***	***

Note: The statistics expressed for each measure on CEFR task level are the median values across all participants.

In the first study, a set of hypotheses were formulated on the direction of the relationship between each of the 11 eye-tracking measures and processing in relation to the CEFR-target level of the tasks (Brunfaut & McCray, 2015, p. 19). Table 8 reproduces these hypotheses and gives an indication of whether they can be supported on the basis of the new B1 *Opinion Matching* tasks in relation to the first study's results on the A1-, A2-, and B2-target level tasks. A tick signifies that the hypothesis was fully supported; a cross means that support was not found; and both a tick and a cross indicates that there was limited support for the hypothesis.

Table 8: 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 directly is due to the fact that the texts for the harder tasks are longer. Longer texts are required in order to 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).	✓ x
	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.	✓ x
	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.	x
Task processing	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.	✓ x
	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.	x
	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.	x
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.	✓ x
	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.	✓ x

The support for the hypotheses on the global processing measures *Total number of fixations* and *Total fixation time on text and responses* show that the B1 *Opinion Matching* tasks follow the expectation that more complex tasks (i.e., higher CEFR-level target) elicit more processing in a much more pronounced manner than the B1 banked gap-fill tasks. Intuitively, the tendency is logical due to the lengthier texts used at the higher CEFR levels of the test. However, whereas the B1 banked gap-fill tasks' results on these measures were quite similar to the A2-target tasks and there was a vast difference with the B2-target level tasks, the processing on the B1 *Opinion Matching* tasks show significantly different results to the A2-target level tasks on the two global processing measures and, overall, the results show a pattern that is more in line with the hypotheses. Similar conclusions can be drawn for the *Number of forwards saccades* and *Number of regressions* on the text of the B1 *Opinion Matching* tasks as compared to the eye-tracking results of the Brunfaut & McCray (2015) study.

The *Median length of forward saccades* on the texts of the B1 *Opinion Matching* tasks was significantly longer than on the other Aptis reading tasks. Potentially, this relates to the nature of this specific task type which consists of four individual text excerpts, which, although on the same topic, are not continuous prose, rather standalone passages. At the same time, the test-takers need to answer seven items in relation to only four texts. As evidenced in the eye trace videos, test-takers often 'visited' one or all texts more than once, potentially using more expeditious reading when searching for answers to later items or after initial reading.<sup>7</sup> This additional use of expeditious reading approaches, which followed the initial careful reading (or preceded it), may be associated with longer forward eye movements.

In the first study with the B1 banked gap-fill texts, a greater *Proportion of regressive movements* was found on the A1 and B1 than the A2 and B2 texts, which suggested more local parsing and careful (re)processing of the texts on the A1 and B1 gap-fill tasks. The eye-trace videos had also indicated that participants concentrated on local reading of the gaps and the words immediately surrounding the gaps – going back and forth. While completing the B1 *Opinion Matching* tasks, however, this pattern was not observed and proportionally fewer regressions were made than on the old B1 texts, suggesting less local and more global reading.

Similar to the first study, the original hypotheses for the *Median regression length*, and for the *Median fixation duration* and *Sum of fixation time on text per word* were violated. However, as discussed in Brunfaut & McCray (2015), measures related to the duration of fixations on the text may not so much relate to the CEFR-target level *per se*, but to the reading approach required by the task. Lower total fixation times per word indicate that readers may not process the text as thoroughly or may not do so every time they re-read. In essence, this measure seems to indicate more expeditious and global approaches to text processing. Thus, the results of this follow-up study suggest that the test-takers did more global reading, and also sometimes expeditious reading, on the new B1 tasks as compared to the old B1 tasks. With reference to the *Median fixation duration* and the *Median regression length*, potentially, these results might relate to the nature of the task, which requires test-takers to consider seven items in relation to only four texts. This may have led many test-takers to access the texts more than once<sup>8</sup>, using careful and expeditious reading approaches at different points in the (re-)reading process, which might have balanced out differences in these median measures.

The findings provide only partial support for the hypothesis regarding the item processing measure *Number of Aol switches between text and responses*, with test-takers making fewer switches between the text and the items of the B1 *Opinion Matching* tasks as compared to the A1, old B1, and B2 tasks. Proportionally, the test-takers also spent a more limited amount of time looking at the item side of the B1 *Opinion Matching* tasks – as demonstrated through the significant differences between the different CEFR-target level tasks in terms of the measure *Proportion of time spent fixating on responses*.

---

<sup>7</sup> Note that this interpretation is also supported by the stimulated recall data presented in Section 5.3.

<sup>8</sup> This speculation is supported by observations made from revisiting the eye trace videos, and by the stimulated recall data presented in Section 5.3.

A potential explanation might be that the item side of the B1 *Opinion Matching* tasks consists of a list of questions with the same four answer options (i.e. the names of the people associated with each text fragment, which are also printed above each excerpt) and test-takers may have relied on their memory when having processed the item side already.<sup>9</sup>

To summarise, the analyses of 11 eye-tracking measures show that eye movements on the B1 banked gap-fill items (old B1 items) and the B1 *Opinion Matching* items (new B1 items) differ statistically significantly on 9 of the 11 measures, which suggests substantial differences in the cognitive processes associated with these two sets of items. Whereas the eye traces on the B1 banked gap-fill items suggested a largely local careful reading approach, the eye traces on the B1 *Opinion Matching* items support suggestions of a more global reading approach.

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

A further aim of the study was to establish whether there are any differences in test-takers' cognitive processing while completing the B1 *Opinion Matching* tasks depending on their L2 reading proficiency (RQ1b) or their overall L2 proficiency (RQ1c). To this end, measures of (reading) proficiency were obtained through test-takers' performances on the full Aptis test, and their eye-movements were analysed according to their English reading proficiency and overall English language proficiency (as determined by their Aptis test scores) through Spearman's rank-order correlations.

Similar to the Brunfaut & McCray (2015) study, the results of these analyses of eye-movements as a function of test-taker L2 (reading) ability uncovered very few significant relationships. Table 9 presents the three measures which showed statistically significant relationships (at the 0.05 level; highlighted in grey) with test-takers' overall L2 proficiency: *Total number of fixations*, *Number of regressions*, and *Proportion of regressive movements*. No significant associations were found between the eye-movement data and test-takers' L2 reading proficiency.

Table 9: Results eye-tracker analyses in relation to L2 reading proficiency (RQ1b) and overall L2 proficiency (RQ1c)

			Total number of fixations	Number of regressions	Proportion of regressive movements
New B1	L2 reading proficiency	Coefficient	-0.261	-0.241	-0.293
		P-Value	0.207	0.246	0.156
	Overall L2 proficiency	Coefficient	-0.423	-0.486	-0.422
		P-Value	0.035	0.014	0.036

<sup>9</sup> This interpretation is supported by the eye-tracking videos and stimulated recalls which indicate that many test-takers started off the task by reading the questions before starting to read the texts.

The statistically significant relationships between L2 proficiency and the eye-tracking metrics (see Table 9) are all in a negative direction, which indicates that the measures diminished as the L2 proficiency of the test-taker was higher, and which might be interpreted as exemplifying different facets of the greater efficiency of processing by the better performing test takers. More specifically, it was found that participants' L2 proficiency negatively correlated with their total number of fixations, or, in other words, the more proficient test-takers fixated fewer times on the B1 *Opinion Matching* tasks' texts and responses while completing the tasks. Negative correlations were also found between participants' overall L2 proficiency and their number and proportion of regressive movements while reading the texts. Thus, the more proficient test-takers made fewer, in the absolute sense, and also proportionally fewer, backward saccades. Taken together, the results indicate that the difference between the lower- and higher-level test-takers was their ability to process the information in the text the first time they read it, rather than having to re-read and thus increase the total number of fixations and regressions.

In Brunfaut & McCray (2015, p. 20), a number of hypotheses were put forward on test-takers' processing – as demonstrated through their eye movements – in relation to their L2 (reading) proficiency. These hypotheses are repeated in Table 10. Similar to the first study, the above findings on the B1 *Opinion Matching* tasks regarding the relationship between the 11 eye-tracking measures and test-takers' L2 (reading) proficiency can be plotted against these hypotheses. A tick indicates that the hypothesis was fully supported, whereas a cross shows that no support was found.

Overall, on the basis of participants' eye movements, limited processing differences were found according to the test-takers' L2 (reading) proficiency (RQ1b and RQ1c). It should be noted, however, that the relatively small sample size in this study may have masked some smaller yet extant effect sizes, but it should also be taken into consideration that most participants in the study had high levels of (reading) proficiency (see Table 4). Thus, the limited number of relationships found may at least partly reflect the nature of the participant group. Potentially, more hypotheses would be confirmed with a participant group comprising a much wider range of abilities.

Table 10: 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	✓
	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	✓
	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

Measure	Hypothesis with reference to RQ1b & RQ1c	RQ1b Met?	RQ1c Met?
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	✓
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 Aol switches between text and response options	x	x
	Proportion of time spent fixating on response options	x	x

## 5.3 Stimulated recall

Insights into test-takers' cognitive processing while completing the B1 *Opinion Matching* tasks was additionally gained with a second set of data, namely, stimulated recalls produced immediately after completing the tasks.

### 5.3.1. Stimulated recall findings on cognitive processes during Aptis B1 Opinion Matching reading task completion (RQ1)

The overarching research question (RQ1) was: *What cognitive processes do test-takers employ during Aptis B1 Opinion Matching reading task completion?* To shed light on this, stimulated recalls with eye-movement recordings as the stimulus were conducted with the 25 participants on two versions of the task (14 items in total). The test-takers provided the correct answer in 88% of the cases (309 cases), but were unsuccessful in 12% of the cases (41 cases). An overview of the cognitive processes the test-takers used while completing these tasks, as evidenced in their stimulated recalls, is provided in Table 11. A distinction is made between processes reported for correctly-answered items versus incorrectly-answered items. To enable this comparison, percentages are provided. For example, for the 309 times that test-takers gave a correct answer, they reported 297 times using a global careful reading approach, i.e. for 96% of the 309 cases. For the 41 times they gave an incorrect answer, they reported 21 times using a global careful reading approach, i.e. for 51% of the 41 cases. In addition, when reading Table 11, it should be kept in mind that often the participants reported using more than one kind of process to establish the answer to a particular item.

Unfortunately, due to confidentiality restrictions and the live status of the tasks used in the study, no direct quotes from the verbal reports can be provided, apart from when it concerns more general comments that do not reveal task content.

Table 11: Stimulated recall results on cognitive processes during Aptis B1 Opinion Matching reading task completion (RQ1)

	Processes	Frequency Item correct (n=309; 100%)		Frequency Item incorrect (n=41; 100%)	
		No.	%	No.	%
Goal setting	Careful reading – global	297	96%	21	51%
	Careful reading – local	76	25%	11	27%
	Expeditious reading – skimming	11	4%	0	0%
	Expeditious reading – search reading	67	22%	6	15%
	Expeditious reading – scanning	21	7%	5	12%
Central core	Creating intertextual representation	50	16%	5	12%
	Creating text level representation	31	10%	2	5%
	Building a mental model	132	43%	10	25%
	Inferencing	280	91%	30	76%
	Establishing propositional meaning	181	59%	16	42%
	Syntactic parsing	8	3%	0	0%
	Lexical access	69	22%	1	5%
	Word recognition	6	2%	6	15%
Additional codes	Background knowledge	0	0%	0	0%
	Collocation	0	0%	0	0%
	Pure guess	0	0%	2	5%
	Test-taking strategies	5	2%	0	0%
	Reading instruction info	20	6%	0	0%
	Pre-reading items	37	12%	0	0%

The stimulated recall findings, presented in Table 11, show that the test-takers used many different cognitive processes while completing the B1 *Opinion Matching* tasks. Overall, when they managed to solve the task correctly, the test-takers had primarily adopted a global, careful reading approach (96%), and also done some expeditious reading in some cases (e.g., search reading, 22%). Interestingly, some test-takers explicitly mentioned first reading the information given in the instruction with the specific goal of gaining global information on the texts (6%). For example<sup>10</sup>,

Participant 5

*I began by reading the instructions. I understood that these people talked about ... [topic].*

Participant 8

*I read the instruction first to know what people were interviewed [about].*

Participant 19

*First I read the instructions. I [then] knew that the passages were about ... [topic] in a ... [text source/type].*

<sup>10</sup> Note that the participants produced their verbal reports primarily in their L1. The quotes given in this report are translations into English of the original L1 reports.

Some test-takers also stated purposefully reading the items (the seven questions) before reading the texts. For example,

Participant 1

*Then I read the questions here [pointing to the item questions on the screen] in order to have rough ideas what these paragraphs [pointing to the text excerpts on the screen] would be about.*

Participant 3

*I was reading the questions in order to know what the passages were going to be about. I knew that question 1 was about ... [question focus]. I learnt from completing the sample items that it was a good idea to start from reading all questions and then reading passages.*

*I read the questions on the left. I thought I could answer them more quickly if I had key words in mind while reading.*

Participant 16

*I started by reading the instructions and the questions one by one. I was thinking that if I could find some links between the questions and the text I was reading, I could answer the questions rights away.*

Participant 19

*I then looked at the questions to see what they were about. This helped me get some ideas about the texts.*

For those items the test-takers answered correctly, they reported using some lower-level processes, particularly establishing propositional meaning (59%) and lexical access (22%), and many higher-level processes. In fact, to arrive at the right answer, the test-takers used inferencing (91%) for most of the items. In addition, they also often built a mental model of different pieces of information within a text (43%), and some of the test-takers established the meaning of the text as a whole (10%) and of different texts in relation to one another (16%) as part of solving some of the items.

It is important to note, however, that in many cases the combined use of processes could be observed from the test-takers' verbal reports. A first example is that, in a number of cases, the reported process involved lexical access and/or establishing meaning at the clause or sentence level, and then inferencing on the basis of this understanding. A second example is that the test-takers demonstrated in their recalls that they had pieced together information from different parts of one of the four texts to form a mental model and inferred on the basis of this. Another example is that they created a textual representation or an intertextual representation (whereby they reasoned through the meaning, differences and similarities of several or all four of the passages), while also making inferences to establish the correct response option.

The most frequently reported interactive combined uses of processes are visualised in Figure 7.

Figure 7: Reported combined uses of processes during B1 Opinion Matching completion

Processes	Frequency Item correct (n=309; 100%)	
	No.	%
 Central core	Creating intertextual representation	50 16%
	Creating text level representation	31 10%
	Building a mental model	132 43%
	Inferencing	280 91%
	Establishing propositional meaning	181 59%
	Syntactic parsing	8 3%
	Lexical access	69 22%
	Word recognition	6 2%

In a very limited number of cases (2%), test-takers reported using test-taking strategy reasoning as part of determining the correct answer, usually in addition to forming an understanding of the text. For example, Participant 22 said: *Firstly, [Text X] hasn't been selected yet to any of the questions. Secondly, [content of Text X, and comparing content of Text X to Text Y].*

For those items which the test-takers had not managed to solve correctly, they had made use of a similar range of processes (see Table 11), but proportionally reported less processing use, and, given the incorrect answers, the test-takers had used these processes in an ineffective manner. Also, when a test-taker had made a pure guess, this did not lead to the correct answer.

To summarise, based on the stimulated recall findings, the B1 *Opinion Matching* reading tasks elicit the spectrum of cognitive and metacognitive processes incorporated in the Khalifa & Weir (2009) model. Particularly extensive usage was reported of careful global reading approaches, of meaning-making at sentence/clause level in combination with inferencing, and of piecing various sentences of a text together to then make inferences. In addition, in a number of cases, some test-takers focused on representing a passage as a whole and/or making links between several passages.

Further insights into test-takers' cognitive processing during B1 *Opinion Matching* task completion were gained through three subanalyses (in line with the first study, and in order to answer the sub-RQs). The findings of these in-depth analyses are presented in the sections below. It should be noted that the focus of these is on the correctly answered items only, since the cognitive processing involved in completing such items is vital information for test validation research.

### 5.3.2 Stimulated recall findings on differences in cognitive processes between the 'old' and 'new' B1 task type

The first subquestion (RQ1a) targeted a comparison between the 'old' and the 'new' B1 tasks designed for the Aptis reading test and asked: *Are there any differences in cognitive processes depending on the task type, i.e. the new B1 Opinion Matching tasks versus the original B1 banked gap-fill tasks?*

Table 12 gives an overview of the cognitive processes the test-takers reported using when successfully completing the 'old' B1 items (banked gap-fill; Brunfaut & McCray, 2015, p. 39) and the 'new' B1 items (matching opinions). To gauge similarities and differences between the two tasks, the reader is referred to the columns presenting the data in percentages (proportion of times test-takers reported using a particular process when giving the correct answer to the 'old' B1 items versus proportion of times test-takers reported using a particular process when giving the correct answer to the 'new' B1 items).

Table 12: Stimulated recall results on cognitive processes when correctly completing B1 banked gap-fill versus B1 Opinion Matching items (RQ1a)

	Processes	Old B1 items (n=279; 100%)		New B1 items (n=309; 100%)	
		No.	%	No.	%
Goal setting	Careful reading – global	60	22%	297	96%
	Careful reading – local	248	89%	76	25%
	Expeditious reading – skimming	7	3%	11	4%
	Expeditious reading – search reading	0	0%	67	22%
	Expeditious reading – scanning	0	0%	21	7%
Central core	Creating intertextual representation	0	0%	50	16%
	Creating text level representation	7	3%	31	10%
	Building a mental model	47	17%	132	43%
	Inferencing	38	14%	280	91%
	Establishing propositional meaning	134	48%	181	59%
	Syntactic parsing	114	41%	8	3%
	Lexical access	56	20%	69	22%
	Word recognition	0	0%	6	2%
Additional codes	Creating paragraph level representation	0	0%	n/a	n/a
	Background knowledge	17	6%	0	0%
	Collocation	58	21%	0	0%
	Pure guess	2	0.7%	0	0%
	Test-taking strategies	n/a	n/a	5	2%
	Reading instruction info	n/a	n/a	20	6%
	Pre-reading item questions	n/a	n/a	37	12%

As can be observed from Table 12, the old and the new B1 items elicited vastly different reading approaches for the majority of the cases, and also noticeable differences in cognitive processing. Whereas the participants in this study had reported primarily using a careful, global reading approach (96%) and some expeditious reading to successfully solve the B1 *Opinion Matching* items, Brunfaut & McCray (2015) observed mainly careful, local reading approaches to the B1 banked gap-fill items (89%).

In addition, although the participants in each study reported using a range of cognitive processes – including both lower- and higher-level reading processes – the use of lower-level processes was proportionally more often reported for the B1 banked gap-fill items (lexical access – syntactic parsing – establishing propositional meaning). Although lower-level processes (in particular, establishing propositional meaning) were also evidenced in the stimulated recalls on the B1 *Opinion Matching* items, proportionally more higher-level processes were observed in the recalls of these 'new' items – in particular inferencing, building a mental model, and also textual and intertextual representation. Furthermore, another clear difference is that, based on the stimulated recalls, successfully solving the B1 *Opinion Matching* items does not depend on collocational knowledge, whereas explicit use of such knowledge was observed with reference to several B1 banked gap-fill items.

Based on the findings on both tasks types targeting the same CEFR level, and recalls from similar populations in terms of background characteristics and English proficiency, it can be concluded that the B1 banked gap-fill and *Opinion Matching* items designed for the Aptis reading test involve at least partly different forms, balances or combinations of processing to correctly solve the items. Whereas test-takers demonstrated a lot of local reading and careful parsing of information around the gaps in the B1 banked gap-fill items, the test-takers completing the B1 *Opinion Matching* items indicated most often to make meaning at and beyond the sentence or text level in combination with extensive use of inferencing from their textual understandings.

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

The second and third sub-questions aimed to explore potential differences in cognitive processing while completing B1 *Opinion Matching* items depending on the test-takers' English reading proficiency (RQ1b) and their overall English proficiency (RQ1c). Participants' English (reading) proficiency was established by means of the full Aptis test.

#### Cognitive processes depending on test-takers' L2 reading proficiency

In order to answer RQ1b – *Are there any differences in cognitive processes during B1 Opinion Matching task completion depending on test-takers' L2 reading proficiency, as measured by the Aptis reading component?* – the stimulated recall data for the successfully completed items were analysed separately according to the CEFR reading proficiency level of the participants. In practice, only one test-taker's reading proficiency was assessed as at the A2 level, four at the B2 level, and 20 in the C range.<sup>11</sup> For the purposes of this analysis, it was not considered meaningful or representative to look into the data of only one test-taker at A2 level, and thus it was decided to only focus on the recall data from participants evaluated to be at B2 or C for English reading.

Table 13 shows the cognitive processes of these 24 test-takers while completing the B1 *Opinion Matching* items, as evidenced in their stimulated recalls, and split according to their reading proficiency. Apart from reporting the raw data per reading proficiency group, the mean (M) number of exhibited cognitive process use per person is also provided to enable proportion interpretations, and a corrected mean is given for the B2 group due to differences in the proportion of correctly answered items between the two reading proficiency groups. Because of the small number of participants per group (in particular the B2 group), no comparative statistical tests were run.

#### B2 level readers

As demonstrated in the stimulated recalls, the B2 readers reported employing a number of lower- and higher-level cognitive processes during completion of the B1 *Opinion Matching* items. On average, they had mostly used a careful, global reading approach (M=12.00). They also most often used inferencing processes (M=12.00), established the meaning of clauses or sentences (M=8.75), and built a mental model of various parts of the texts (M=3.75) while successfully completing the items.

---

<sup>11</sup> Note that the Aptis score reporting system does not distinguish between the C1 and C2 levels of the CEFR.

### C level readers

The C readers similarly reported many different lower- and higher-level cognitive processes while completing the B1 *Opinion Matching* items, and, on average, mostly adopted a careful, global reading approach ( $M=11.90$ ). Also, they most often used inferencing processes ( $M=10.75$ ), established the meaning of clauses or sentences ( $M=6.75$ ), and built a mental model of various parts of the texts ( $M=5.56$ ) while successfully completing the items.

### The two reading proficiency levels

The results descriptions and the data presented in Table 13 show that both groups of participants had mostly targeted gaining global comprehension while carefully reading the texts. Both groups also evidenced using several different types of cognitive processes while completing the items, including lower- and higher-level processes (even intertextual representation in some cases) and with a large role for inferencing processes. Some minor differences could be detected in the use of the lower-level process 'establishing propositional meaning' ( $M_{\text{corrected}_{B2}}=9.01$ ,  $M_C=6.75$ ) which was on average more employed by the B2 readers, and the higher-level processes 'building a mental model' ( $M_{\text{corrected}_{B2}}=3.86$ ,  $M_C=5.56$ ), and 'creating intertextual representation' ( $M_{\text{corrected}_{B2}}=0.52$ ,  $M_C=2.25$ ) which were on average more used by the C-range readers.

Table 13: Stimulated recall results on cognitive processes of correct items depending on test-takers' L2 reading proficiency (RQ1b)

	Processes	B2 reading proficiency (n=4)			C reading proficiency (n=20)	
		No.	M	M corrected	No.	M
Goal setting	Careful reading – global	48	12.00	12.36	238	11.90
	Careful reading – local	15	3.75	3.86	56	2.80
	Expeditious reading – skimming	4	1.00	1.03	7	0.35
	Expeditious reading – search reading	14	3.50	3.61	44	2.20
	Expeditious reading – scanning	6	1.50	1.55	15	0.75
Central core	Creating intertextual representation	2	0.50	0.52	45	2.25
	Creating text level representation	3	0.75	0.77	26	1.30
	Building a mental model	15	3.75	3.86	113	5.56
	Inferencing	48	12.00	12.36	215	10.75
	Establishing propositional meaning	35	8.75	9.01	135	6.75
	Syntactic parsing	3	0.75	0.77	4	0.20
	Lexical access	10	2.50	2.58	53	2.65
	Word recognition	0	0.00	0.00	3	0.15
Additional codes	Background knowledge	0	0.00	0.00	0	0.00
	Collocation	0	0.00	0.00	0	0.00
	Pure guess	0	0.00	0.00	0	0.00
	Test-taking strategies	0	0.00	0.00	5	0.25
	Reading instruction info	4	1.00	1.03	14	0.70
	Pre-reading item questions	6	1.50	1.55	29	1.45
	<b>Total</b>	<b>213</b>	<b>53.25</b>	<b>54.85</b>	<b>1002</b>	<b>50.10</b>

To summarise, both CEFR B2 and C-level readers reported using a wide range of reading processes while solving B1 *Opinion Matching* tasks and reported using similar processes to similar extents. Slight differences were observed in terms of B2 readers' somewhat more extensive reliance on meaning-making at the clause or sentence level, and C-range readers' building of mental models of the text and seeking connections between texts.

It should be kept in mind, however, that the number of participants was small, in particular for the B2 group, and that all participants had relatively high levels of reading proficiency. Thus, these conclusions on lack of substantial processing differences depending on L2 reading proficiency have to be treated with care and might not be generalisable to less proficient readers.

### **Cognitive processes depending on test-takers' overall L2 proficiency**

The third subquestion of the study (RQ1c) was: *Are there any differences in cognitive processes during B1 Opinion Matching task completion depending on test-takers' overall L2 proficiency, as measured by all different Aptis components?* To investigate this, the stimulated recall data of the successfully completed B1 *Opinion Matching* items were analysed separately according to the participants' overall English proficiency.

Two proficiency level groups were established on the basis of participants' total scores on the full Aptis test. The 13 lowest-scoring participants formed the 'lower proficiency half' (M=160.2, SD=8.0), the 12 highest-scoring participants constituted the 'higher proficiency half' (M=180.8, SD=6.9).

The two proficiency groups' cognitive processes while completing the B1 *Opinion Matching* items, as verbalised during the stimulated recalls, are provided in Table 14. Raw data, as well as means (M) for cognitive process use per person, are included to allow for an easier proportion interpretation. A corrected mean is also given for the lower-proficiency group, because of proportional differences in correctly answered items and in order to allow for comparisons between the two proficiency groups. Again, no comparative statistical tests were run due to the relatively small numbers per group.

#### **Lower proficiency half**

The stimulated recall data of the lower proficiency group indicate that they mostly adopted a careful, global reading approach (M=11.00) to solving the B1 *Opinion Matching* items and that they used various lower- and higher-level cognitive processes to successfully complete these items. On average, they most often made inferences (M=12.00), focused on understanding clauses and sentences (M=7.15), or pieced information together from various parts of a text (M=4.92).

#### **Higher proficiency half**

Participants in the higher proficiency group similarly mostly conducted careful reading and global reading processes (M=12.83), and used a range of lower- and higher-level cognitive processes to determine the correct answers to the B1 *Opinion Matching* tasks. Particularly, they often inferred from the texts (M=10.33), established propositional meaning (M=7.33), and built a mental model (M=5.67).

#### **The two proficiency groups**

Both proficiency groups were observed to use a wide spectrum of cognitive processes while completing B1 *Opinion Matching* items, including lower- and higher-level processes. Proportionally, and on average, both groups reported using different processes to the same extent; the processing tendencies are quite similar between the lower- and higher proficiency groups.

In sum, no substantial processing differences were found for successful completion of the B1 *Opinion Matching* items between two groups of participants split according to overall English proficiency. It should be noted, however, that the present study's population (as was the case in the first study) overall had a (high-)intermediate to advanced level proficiency profile; more distinct differences might be found for more diverse L2 proficiency populations.

Table 14: Stimulated recall results on cognitive processes of correct items depending on test-takers' L2 proficiency (RQ1c)

	Processes	Lower proficiency half (n=13)			Higher proficiency half (n=12)	
		No.	M	M corrected	No.	M
Goal setting	Careful reading – global	143	11.00	11.24	154	12.83
	Careful reading – local	49	3.77	3.85	27	2.25
	Expeditious reading – skimming	5	0.38	0.39	6	0.50
	Expeditious reading – search reading	38	2.92	2.99	29	2.42
	Expeditious reading – scanning	13	1.00	1.02	8	0.67
Central core	Creating intertextual representation	20	1.54	1.57	30	2.50
	Creating text level representation	11	0.85	0.86	20	1.76
	Building a mental model	64	4.92	5.03	68	5.67
	Inferencing	156	12.00	12.26	124	10.33
	Establishing propositional meaning	93	7.15	7.31	88	7.33
	Syntactic parsing	6	0.46	0.47	2	0.17
	Lexical access	34	2.62	2.67	35	2.92
Additional codes	Word recognition	5	0.38	0.39	1	0.08
	Background knowledge	0	0	0.00	0	0.00
	Collocation	0	0	0.00	0	0.00
	Pure guess	0	0	0.00	0	0.00
	Test-taking strategies	1	0.08	0.08	4	0.33
	Reading instruction info	8	0.62	0.63	12	1.00
	Pre-reading item questions	19	1.46	1.49	18	1.50
	<b>Total</b>	<b>665</b>	<b>51.15</b>	<b>52.28</b>	<b>626</b>	<b>52.17</b>

## 6. DISCUSSION

The findings of both the eye-tracking and stimulated recall analyses indicate that the B1 *Opinion Matching* tasks designed for the Aptis reading test elicit a wide range of cognitive processes from test-takers on correctly-answered items. Evidence was found for the use of the various metacognitive and cognitive processes theorised in Khalifa & Weir's (2009) model of reading. No obvious risks for construct-irrelevant variance were identified in the dataset (word spotting/matching or guessing did not lead to correct answers, and very little test-taking strategy use was reported).

Based on the Aptis General Technical Manual, Version 1.0 (O'Sullivan & Dunlea, 2015, p. 13), the aim of the B1-target level tasks is to assess "[t]ext-level comprehension of short texts" through whole-text reading, requiring "[c]areful global reading" involving "text-level comprehension and reading beyond the sentence[-level]". In this respect, Brunfaut & McCray (2015) had found that the majority of cognitive processes the test-takers used to successfully complete the original B1 banked gap-fill tasks were careful, local reading and sentence-level processing, and that there was a risk of construct irrelevant-variance through use of background and collocational knowledge.

Thus, there was a discrepancy between what the test developers had intended and the cognitive processes elicited by the original task type. In the present study, which explored test-takers' cognitive processing while successfully completing newly designed and trialled B1 *Opinion Matching* tasks, it was found that test-takers did some careful, local reading and expeditious reading, such as search reading, skimming and scanning, but they overwhelmingly adopted a careful, global reading approach aiming to understand the texts as a whole. While completing the tasks, the test-takers also used lower-level processes, such as lexical processing and meaning-making at clause or sentence-level, but they additionally often made use of higher-level processes whereby they focused on meaning-making beyond the sentence-level, at text-level and sometimes also between different texts. Importantly, they extensively made use of inferencing (in combination with other processes) to establish the right answer to the items. The cognitive processing evidenced on the new B1 *Opinion Matching* items thus seems suitably aligned with the intended careful global reading and higher-level processing for the B1 target-level of the Aptis reading test (as described in O'Sullivan & Dunlea, 2015).

## 7. CONCLUSION

---

The aim of this study was to examine test-takers' cognitive processing while responding to B1 *Opinion Matching* items designed for the Aptis reading comprehension test. The study was designed to follow up on the results of an earlier study on cognitive processing during Aptis reading test completion (i.e. Brunfaut & McCray, 2015), in particular in response to discrepancies between intended and actual processing on the B1 banked gap-fill tasks of the Aptis reading test.

In-depth insights into the newly developed B1 *Opinion Matching* tasks were gained by means of eye-tracking during task completion and stimulated recalls immediately after task completion, with eye-movement recordings acting as the stimulus. The processing patterns and tendencies observed in the eye-tracking visualisations, the statistical analyses of various eye-tracking measures, and the analyses of test-takers' verbal reports were mutually confirmatory. The triangulation of the findings provides a solid basis for conclusions on test-takers' cognitive processes and the validity of the B1 *Opinion Matching* tasks.

It was found that the entire range of cognitive processes, as specified by Khalifa & Weir (2009), was used by test-takers while completing B1 *Opinion Matching* tasks. Some evidence was found of expeditious and careful local reading approaches, but the majority of correct items had involved test-takers in careful, global reading. The test-takers also demonstrated making use of lower-level processes (e.g., lexical access and propositional meaning building), but very often also higher-level processes (e.g., building a mental model, creating text level or intertextual representations). In particular, processes such as 'lexical access', 'establishing propositional meaning', 'building a mental model' and 'creating text-level / intertextual representation' were used to gain an overall understanding of (various pieces of information in) the texts and, on the basis of comprehension of this textual information, inferences were made to answer the items.

The findings indicate that the cognitive processing patterns on the B1 *Opinion Matching* tasks are generally in line with the Aptis intended target processes for this level of the reading test.

## REFERENCES

---

- Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.
- Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, 30(4), 441–465.
- Brunfaut, T. & McCray, G. (2015). *Looking into test-takers' cognitive processes whilst completing reading tasks: a mixed-method eye-tracking and stimulated recall study*. ARAGs Research Reports Online, Vol. AR/2015/001. London: The British Council.  
[https://www.britishcouncil.org/sites/default/files/brunfaut\\_and\\_mccray\\_report\\_final.pdf](https://www.britishcouncil.org/sites/default/files/brunfaut_and_mccray_report_final.pdf)
- Cohen, J. (1988). *Statistical power analysis for behavioral sciences*. Hillsdale, New Jersey: Erlbaum.
- Council of Europe (2001). *Common European Framework of Reference for languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
- Holmqvist, K., Nyström, M., Anderson, R., Dewhurst, R., Jarodzka, H. & van de Weijer, J. (2011). *Eye tracking*. Oxford: Oxford University Press.
- Israel, S. E. (2015). *Verbal protocols in literacy research: Nature of global reading development*. NY/Oxon: Routledge.
- Khalifa, H. & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge: Cambridge University Press.
- O'Sullivan, B. (2015). *Aptis test development approach*. Technical Report; Vol. TR/2015/001. London: The British Council.
- O'Sullivan, B. & Dunlea, J. (2015). *Aptis General Technical Manual Version 1.0*. Technical Report TR/2015/005. British Council: London.
- Pressley, M. & Afflerbach, P. (1995). *Verbal protocols of reading*. New Jersey: Lawrence Erlbaum.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372–422.

## British Council Assessment Research Group

The Assessment Research Group was formed in 2013 to support the British Council's work in assessment and testing across the world. The team is responsible for ensuring that all new assessment products and new uses of existing products are supported by the most up-to-date research. They also continuously evaluate the quality of British Council assessment products.

### LOOKING INTO READING II: A FOLLOW-UP STUDY ON TEST-TAKERS' COGNITIVE PROCESSES WHILE COMPLETING APTIS B1 READING TASKS

VS/2016/001

**Brunfaut**

BRITISH COUNCIL VALIDATION  
SERIES

Published by British Council  
10 Spring Gardens  
London SW1A 2BN

© **British Council 2016**

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

[www.britishcouncil.org/aptis/research](http://www.britishcouncil.org/aptis/research)