BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

LOOKING INTO LISTENING:
Using eye-tracking to establish the cognitive validity of the Aptis Listening Test

AR-G/2017/3

F. Holzknecht and K. Eberharter (joint first authors)
B. Kremmel, M. Zehentner, G. McCray, E. Konrad and C. Spöttl

ARAGs RESEARCH REPORTS ONLINE
SERIES EDITOR: VIVIEN BERRY
ISSN 2057-5203

ABSTRACT

This study investigated the cognitive processing of 30 test-takers while completing the Aptis Listening Test. The research studied test-takers' processes according to ones targeted at the different item levels in the Aptis Test.

Specifically, it examined whether test-takers' cognitive processes and types of information used corresponded to the ones targeted at the different CEFR levels. To this end, a detailed analysis of test-takers' verbal recalls was conducted, which were stimulated by a replay of their eye-traces while they had been solving the items. The study also explored the usefulness of quantitative analyses of eye-tracking metrics captured during listening tests.

The stimulated recall findings indicate that the Aptis Listening Test successfully taps into the range of cognitive processes and types of information intended by the test developers. The data also shows, however, that the differences between the CEFR levels in relation to the intended cognitive processes could be more pronounced, and that the process of "discourse construction" could be more evident for B2 items. It is, therefore, suggested that a different item type could help elicit this type of higher-order processing. In terms of types of information used by candidates, a clear difference and progression between the CEFR levels to answer items correctly was observed.

The quantitative analysis of the eye-tracking metrics revealed interesting results. A linear mixed effects model analysis, with visit duration on response options as the dependant variable, showed that test-takers looked at the response options of higher-level items significantly longer than at the response options of lower-level items. The results also showed that response options higher up on the screen were looked at significantly longer than response options lower down, regardless of item level. In addition, it was found that better readers focused on the response options significantly longer than poorer readers.

Authors

Franz Holzknecht holds an MA in Language Testing from Lancaster University and an MA in English Language Teaching and Sports Education from the University of Innsbruck. He works as a researcher for the Language Testing Research Group at the University of Innsbruck and is also pursuing his PhD on listening assessment at Lancaster University. His research interests are mainly in language assessment, particularly in listening assessment, writing assessment, and the use of eye-tracking. Franz was professionally involved in language test development for the standardised Austrian school-leaving examination. He has taught language testing courses at teacher training colleges in Austria and England, and has counselled a number of international language testing companies on test development issues. Franz has presented his research at international conferences and has published his work in *Papers in Language Testing and Assessment* and in several book chapters.

Kathrin Eberharter attained her first MA at the University of Innsbruck, qualifying her to teach EFL and German in Austrian secondary schools. During her final practical year of teacher training, Kathrin started part-time work for the Austrian exam-reform project at the University of Innsbruck. Kathrin later continued to work full-time as project assistant and test moderator for the Austrian school-leaving examination and, soon after, began her MA studies with Lancaster University's Language Testing MA (by distance) program. She currently holds a pre-doc position at the University of Innsbruck, where she teaches pre-service courses related to language testing and assessment. Kathrin is enrolled in Lancaster University's Applied Linguistics PhD program, and her current interests lie particularly with the processes involved in the rating of speaking and writing.

Benjamin Kremmel holds an MA in Language Testing from Lancaster University and an MA in English Language Teaching and Psychology and Philosophy Education from the University of Innsbruck. He is a researcher and lecturer at the University of Innsbruck, Austria, and is enrolled as a PhD student at the University of Nottingham, UK. His research interests are in language testing, particularly in the areas of vocabulary assessment, L2 reading assessment and language assessment literacy. Benjamin was involved in the Austrian SRDP project, which developed standardised national school-leaving exams for the modern languages. He is the recipient of the *2013 Caroline Clapham IELTS Masters Award* and the winner of the *Robert Lado Memorial Award 2015*. Benjamin has presented his work at numerous international conferences. His research has been published in *Language Testing, Language Assessment Quarterly, Applied Linguistics*, *TESOL Quarterly,* and *Papers in Language Testing and Assessment*.

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

Matthias Zehentner holds an MA in English Language Teaching and Mathematics Education from the University of Innsbruck and is enrolled in the distance MA program in Language Testing at Lancaster University. He works as a researcher at the University of Innsbruck and teaches English and Mathematics in a secondary technical vocational school in Innsbruck, Austria. His research interests focus on language assessment, particularly on assessing the productive skills, factors affecting spoken performance, task development, and rater behaviour. Matthias was involved in the development of standardised tasks for the Austrian school-leaving exam reform project (SRDP) and contributed to the construction of rating scales for the oral part of the exam. Matthias has presented his work at national and international conferences.

Eva Konrad holds an MA in Language Testing from the University of Lancaster and an MA in English and American Studies from the University of Innsbruck. She works as a research assistant at the University of Innsbruck where she is also a member of the Language Testing Research Group. Eva was previously involved in the development of language tests for the standardised Austrian school leaving exam project, and she has lectured at Austrian teacher training colleges. For her Masters thesis, she conducted research into the potential effects of using a dictionary in a writing exam. Eva is currently enrolled in Lancaster's PhD program in Language Testing, where she plans to investigate diagnostic testing of L2 listening. Her research interests are the assessment of L2 writing and listening, diagnostic testing and reading and listening task development.

Carol Spöttl is the co-ordinator of the Language Testing Research Group (LTRGI) at the University of Innsbruck's School of Education. She has degrees from the Universities of Edinburgh, East Anglia and Lancaster, and has worked at the University of Innsbruck in the field of language teaching, testing and evaluation. From 2007- to 2015, Carol was the exam reform project leader in Austria. The government-funded project introduced a new school-leaving exam in the foreign languages. It saw the development of CEFR linked tests at two CEFR levels, for the skills reading, listening, writing and language in use and for four of the languages taught: English, French, Italian and Spanish. Current research projects in which the LTRGI is involved include eye-tracking studies for listening tests, identifying anchor items, providing teacher trainers with benchmarked performances for speaking and writing, and comparing writing assessment in two European countries to cross-disciplinary projects with the university's medical faculty.

CONTENTS

1. BACKGROUND 6

2. LISTENING COGNITION RESEARCH METHODOLOGY 7

2.1 Eye-tracking 8
2.2 Stimulated recall 8

3. RESEARCH QUESTIONS 9

4. METHODOLOGY 9

4.1 Pilot study 9
4.2 Main study 11
4.2.1 Participants 11
4.2.2 Materials 12
4.2.3 Procedure 12
4.2.4 Ethical consent 14
4.2.5 Data analysis 14

5. RESULTS 19

5.1 Descriptive statistics 19
5.1.1 Aptis Listening Test and measure of receptive English proficiency 19
5.1.2 Listening once vs. listening twice in the Aptis Listening Test 20
5.2 Eye-tracking 20
5.2.1 Linear mixed model 20
5.3 Stimulated recall 23
5.3.1 Cognitive processes used to answer items on the Aptis Listening Test (RQ1.1) 24
5.3.2 Cognitive processes at the different item levels of the Aptis Listening Test (RQ 1.2) 25
5.3.3 Types of information used to answer items on the Aptis Listening Test (RQ2.1) 27
5.3.4 Types of information at the different item levels of the Aptis Listening Test (RQ2.2) 28

6. DISCUSSION AND CONCLUSION 30

6.1 Cognitive processes and types of information used 30
6.2 Unexpected findings: response order and reading ability 31
6.3 Potential and limitations of the methodology 31
6.4 Areas for future research 32

REFERENCES 33

Figures

Figure 1: Histogram of visit duration on response options 21
Figure 2: Transformed distribution of visit durations on response options 21

Tables

Table 1: APTIS CEFR levels mapping onto the cognitive processing model of listening and the type of information targeted 7
Table 2: Different stimulated recall procedures in the pilot study 10
Table 3: Pilot study design 10
Table 4: Description of dependent and control variables 16
Table 5: Description of exploratory variables 16
Table 6: Coding scheme for coding transcribed stimulated recall data 18
Table 7: Descriptive statistics of the different test packages 19
Table 8: CEFR levels of participants 19
Table 9: CEFR level by number of times listened to the recording for the Aptis Listening Test 20
Table 10: Results of the linear mixed model 21
Table 11: Cognitive processes employed to answer Aptis Listening Test items 24
Table 12: Cognitive processes employed per target CEFR level (overall, in percent of cases) 25
Table 13: Cognitive processes employed per target CEFR level 25
Table 14: p-Values of pairwise comparisons of cognitive process ratios across CEFR levels 27
Table 15: Information used to answer Aptis Listening Test items 28
Table 16: Information used per target CEFR level (overall, in percent of cases) 28
Table 17: Information used per target CEFR level (correct answers only, in percent of cases) 29
Table 18: p-Values of pairwise comparisons of information type ratios across CEFR levels) 30

1. BACKGROUND

Despite the growing number of research studies on the challenges of assessing listening (see for example, Taylor & Geranpayeh, 2013), research into the cognitive processes underlying the listening construct is still sparse. In this respect, Vandergrift's observation of 2007, that despite the fact that "[l]istening comprehension lies at the heart of language learning, […] it is the least understood and least researched skill" (2007, p. 191) still holds true. It can be argued that understanding the cognitive processes listeners employ when trying to make sense of an auditory signal should be the central focus in listening assessment research. Cognitive processing is directly related to construct validity, and as Buck aptly states, "ensuring that the right construct is being measured is the central issue in all assessment" (Buck, 2001). Some recent studies have shown the importance and potential impact of this strand of research (see, for example, Field, 2013). However, the cognitive processes of listening test-takers are still not well understood. This current study addresses this need in the context of the Aptis Test.

The Aptis Test is an online English language assessment tool for adults aged 16 and above, developed by the British Council. It includes individual tests for the four language skills: reading, listening, speaking, and writing; as well as a component assessing grammar and vocabulary. Test-takers can choose which of the four skills they would like to be tested on. The Grammar and Vocabulary Test is taken by all test-takers.

One Aptis Listening component consists of 25 short recordings with separate four-option multiple-choice questions for each recording. The 25 questions are linked to four levels of the Common European Framework (A1, A2, B1 and B2) and they appear in sequence in terms of their difficulty, starting at the lower levels. In addition to their relation to the CEFR, the test specifications of the listening test also include sections on cognitive processing and targeted information (O'Sullivan & Dunlea, 2015, pp. 48–51).

The theoretical validation procedure proclaimed by Aptis is based on the adapted model of Weir's (Weir, 2005) socio-cognitive framework established by O'Sullivan and Weir (2011) and O'Sullivan (2011). This model consists of three main parts: the test-taker, the test system and the scoring system. The research in this paper focuses on two of these parts in relation to the Aptis Listening Test: the test-taker and the test system. In particular, it will investigate whether the cognitive processes employed by the test-takers and the linguistic demands of the recordings for the individual items are in line with the item designers' intentions and the exam's test specifications.

Aptis divides the four difficulty levels of the items in the Listening Test (A1 to B2) according to different levels of cognitive processing, following Field's (Field, 2008, 2013) bottom-up processing model. The model includes both lower-level and higher-level listening processes. According to Field, listeners first identify incoming sound signals as speech and then phonologically decode these sounds ("input decoding"). Following successful decoding, listeners combine the different phonemes to form individual words ("lexical search") and syntactic patterns at clause level ("parsing"). These lower-level processes are then followed by "meaning construction", which is characterised by relating the bare meaning of clauses and sentences to the context. Finally, in the "discourse construction" stage, listeners make sense of the speech event as a whole, by relating everything that has been said so far to the overall meaning of the message.

The developers of the Aptis Test claim that test-takers need to master these processes to be able to answer the questions at the different levels correctly. According to the Aptis test specifications, A1 items target lexical search processes, A2 and B1 items target meaning construction as well as discourse construction processes, and B2 items are testing solely discourse construction processes, as illustrated in Table 1. To date, it has not been systematically established whether test-takers actually employ the processes at the different levels as intended by the test developers.

APTIS CEFR level	Cognitive processes (Field, 2013)	Type of information targeted
B2	Discourse construction	Interpretive meaning at the utterance level, Meaning at the discourse level
A2 / B1	Discourse construction, Meaning construction	Factual information
A1	Lexical search	Lexical recognition

Table 1: APTIS CEFR levels mapping onto the cognitive processing model of listening and the type of information targeted

Directly related to the cognitive processing levels, the Aptis Listening test specifications also outline which kind of information is targeted by the questions at the individual levels. There are four different types of information targeted by the test: lexical recognition, factual information, interpretive meaning at the utterance level, and meaning at the discourse level. The type of information targeted varies from one level to the next, from lexical recognition for A1 items, to factual information for A2 and B1 items, and interpretive meaning at the utterance level and meaning at the discourse level for B2 items, as shown in Table 1 above. Grading the targeted information this way seems logical, as it parallels the increasing complexity of the cognitive processes involved. However, it has not yet been investigated whether test-takers actually make use of these proposed types of information to answer the questions at the different levels, or whether they arrive at the answer using different information.

The present study addresses these research gaps. It investigates which cognitive processes test-takers employ when solving items of the Aptis Listening Test. It also aims to identify which information in the recording and in the test questions test-takers use to arrive at their answers. The study will do so by means of analysis of online eye-movements captured while responding to test items and retrospective stimulated recalls aided by participants reviewing and rationalising their eye-movements, as will be outlined in the following section.

2. LISTENING COGNITION RESEARCH METHODOLOGY

Due to the nature of listening, researchers are somewhat limited in their choice of methods for investigating the thought processes of listening test-takers. One promising research method, which has been increasingly used over the last years in test-taker cognition studies, is eye-tracking. However, listening researchers are constrained in applying this method, as the main mode in listening tests is not visual but oral. Thus, when interpreting eye-tracking data on listening tests, it is not clear whether eye-movements are indicative about listening processing, reading processing, or test-taking processing. Apart from eye-tracking, one of the most commonly used research methods in test-taker cognition studies is introspection through verbal reports (Ericsson & Simon, 1987, 1993), i.e. asking test-takers to vocalise their thought processes while they are engaged in the activity under scrutiny (concurrent verbal reports) or sometime after the activity has been finished (retrospective verbal reports). Although some researchers have pointed out that concurrent verbal reports should be used instead of retrospective verbal reports because the latter might be influenced by memory restraints (see, for example, Green, 1998, p. 6), listening cognition researchers rely on retrospective verbal reports, as test-takers cannot think aloud while they are engaged in the activity. In order to minimise memory effects, subjects can be provided with a stimulus to re-activate their thought processes, and produce a stimulated recall (Gass & Mackey, 2000). In the following, it will be reviewed how these two methods (eye-tracking and stimulated recall) have been used to investigate listening test-takers' thought processes.

2.1 Eye-tracking

Although eye-tracking has been used as research methodology to investigate test-taker processes in reading in a number of studies (see, for example, Bax, 2013; Bax & Weir, 2012; Brunfaut & McCray, 2015; McCray, 2013; McCray, Alderson, & Brunfaut, 2012; McCray & Brunfaut, 2016), its employment to inform listening processing models has been sparse. That is because in contrast to reading, where "it has become increasingly clear that eye-movements provide a very good (and precise) index of mental processing" (Rayner, 2009, p. 1487), the link between eye-movements and cognitive processing in listening is only indirect. Still, studies have shown that eye-tracking methodology can inform certain aspects of listening processing. For example, there is a large body of research on the so-called "visual world paradigm" (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), which proclaims that there is a strong relationship between what listeners hear and what they look at on a screen. Visual world paradigm experiments usually involve a small number of simple visual objects and short audio files related to these objects, such as individual words or sentences. However, cognition research employing eye-tracking on more complex listening tasks typically used in listening assessment is only just emerging.

One of the few studies available using eye-tracking to investigate listening test-takers' thoughts is Winke and Lim (2014). In their study, Winke and Lim found different eye-movement patterns between candidates with high and low test-taking anxiety when taking an IELTS Listening Test, as well as between high-scoring and low-scoring candidates on the same test. For example, for fill-in-the-blank questions, more anxious test-takers spent considerably more time looking at key words than less anxious test-takers, and high-scoring test-takers fixated the key words surrounding the blanks significantly more quickly than low-scoring test-takers. Although Winke and Lim were able to show these relationships through eye-tracking, they highlight that when using this method to investigate test-takers' cognitive processes in listening tests "[i]t is impossible to disentangle the two constructs (reading of the text on the page [and] L2-listening skills) in this context" (Winke & Lim, 2014). Therefore, in order to investigate listening processes more accurately, eye-tracking needs to be used in combination with other research methods, such as stimulated recall.

2.2 Stimulated recall

Stimulated recall (Gass & Mackey, 2000) has been employed in a number of investigations on test-taker cognition in listening assessment (Badger & Yan, 2012; Field, 2012, 2015; Harding, 2011; Winke & Lim, 2014). In the majority of these studies, test-takers' answers were used as stimulus to initiate retrospection, and the authors report that the stimulated recalls produced high quality data. Winke and Lim (2014) followed a slightly different procedure, as their study also involved eye-tracking. Instead of using test-takers' answers as stimulus, they replayed a video of the participants' eye-movements on the last page of the test used in their investigation to initiate test-taker retrospection. Winke and Lim report that a detailed analysis of their stimulated recall data and triangulation with their eye-tracking data is still pending. However, they are hopeful that such an analysis may help them "understand more fully how eye-movement data can be best interpreted when researchers are investigating the complex nature of L2-listening test performance" (Winke & Lim, 2014).

This is in line with research in reading assessment, where it has been shown by a number of studies that the combined use of eye-tracking and stimulated recall gives researchers more confidence in interpreting the cognitive processes of test-takers (Bax, 2013; Bax & Weir, 2012; Brunfaut & McCray, 2015; McCray, 2013; McCray et al., 2012).

3. RESEARCH QUESTIONS

In light of the current state of research, the following research questions will be addressed:

RQ1.1 Which cognitive processes do test-takers use to answer items on the Aptis Listening Test?

RQ1.2 Which cognitive processes do test-takers use at the different item levels (A1, A2, B1 and B2) of the Aptis Listening Test?

RQ2.1 Which types of information do test-takers use to answer items on the Aptis Listening Test?

RQ2.2 Which types of information do test-takers use at the different item levels (A1, A2, B1 and B2) of the Aptis Listening Test?

Based on the literature review, eye-tracking combined with stimulated recall were chosen to investigate these research questions.

4. METHODOLOGY

4.1 Pilot study

Prior to data collection, an extensive pilot study was conducted. The pilot study served a number of purposes. First, it enabled the researchers to test the different stimulated recall procedures identified in the literature review and to determine which of the procedures was the most useful for the study. Second, the layout of the test and the technical integration of the sound files needed to be adapted for use on an eye-tracker. The pilot study was carried out to test the feasibility of the adapted versions. Third, the pilot study was conducted to test the suitability of three different Aptis Listening Test packages available for the study and to choose the most suitable package for the main study. Fourth, during the pilot study, appropriate participants could be sampled both in terms of their general English ability, as well as in terms of their suitability for eye-tracking experiments. Although none of the participants in the pilot study took part in the main study, the recruitment of pilot study participants helped the researchers to find appropriate participants for the main data collection. Finally, the data collected in the pilot study was used to test the different analyses used in the main study, such as the coding of the stimulated recalls, and it enabled the researchers to check the quality of the eye-traces for the eye-tracking data analysis.

Six participants were recruited for the pilot study. In terms of stimulated recall procedure, two different methods were trialled, as identified in the literature review. The underlying procedure was the same for both methods. Each participant performed one of three chosen versions of the Aptis Listening Test, each consisting of 25 items, on a Tobii TX300 eye-tracker. After three to four items, the test was stopped and participants were asked in German to recall in detail how they had arrived at their answer for each item. Participants were free to use German or English for their recalls. In Method 1, the participants' answers served as stimulus for recall (Badger & Yan, 2012; Field, 2012, 2015; Harding, 2011), i.e. during recall the participants were reminded of the answer they gave to each item while simultaneously seeing the item on screen. In Method 2, a recording of the participants' eye traces and mouse movements (Bax, 2013; Bax & Weir, 2012; Brunfaut & McCray, 2015; McCray et al., 2012; Winke & Lim, 2014) overlaid with the sound file for each item was replayed to the participants to aid recall. Table 2 below sums up the two different procedures tested.

Method	Stimulus for recall
Method 1	Answers to the items
Method 2	Recording of participants' eye- and mouse-movements overlaid with the sound file for each item

Table 2: Different stimulated recall procedures in the pilot study

The pilot study followed a fully-crossed design. Each participant performed roughly half of the items (12 or 13 out of 25) for one of the three listening packages following Method 1 and the other half of the items following Method 2. Each of the items for the three different packages was performed by one participant following Method 1 and by another participant following Method 2, as shown in Table 3 below. In addition, after the participants had completed all 25 items, they were asked which of the two stimulated recall methods they found more helpful for remembering their thoughts while answering the items, and to give reasons for their preference.

Participants	Aptis Listening Test version	Stimulated recall method	Items							
1	1	1	1-3		7-9		13-15		19-21	
		2		4-6		10-12		16-18		22-15
2	1	1		4-6		10-12		16-18		22-15
		2	1-3		7-9		13-15		19-21	
3	2	1	1-3		7-9		13-15		19-21	
		2		4-6		10-12		16-18		22-15
4	2	1		4-6		10-12		16-18		22-15
		2	1-3		7-9		13-15		19-21	
5	3	1	1-3		7-9		13-15		19-21	
		2		4-6		10-12		16-18		22-15
6	3	1		4-6		10-12		16-18		22-15
		2	1-3		7-9		13-15		19-21	

Table 3: Pilot study design

The pilot data was analysed in several steps. First, all of the stimulated recalls were transcribed and coded by two researchers, following a coding scheme based on the five levels of cognitive processing by Field and the five different types of information outlined in the Aptis Listening test specifications (see Table 1 above). The coders disagreed in a number of cases and discussed their disagreements to arrive at a final code to be assigned. Cognitive processes generally led to more discussion than types of information. Similarly, higher-level processing levels generally led to more discussion than lower-level processing levels. The coding scheme used was adapted after the pilot study based on the discussions of the two coders. Second, the participants' answers to the follow-up questions on which of the two stimulated recall methods they preferred were transcribed and analysed according to preferred method and reasons for the preference. Finally, the quality of the eye-tracking data was inspected by examining the correspondence between the text location on the screen and the eye-tracking gaze plots produced by the Tobii Studio Pro software, in line with Holmqvist, Nyström, Andersson, Dewhurst, Jarodzka, and Van de Weijer (2011).

The results of the pilot study revealed three important findings. First, the results showed that one of the three Aptis Listening Test versions generated richer responses on a small number of items from the participants during stimulated recall than the other two versions. This was possibly due to the characteristics of the specific items included in this version, such as certain cultural references with which Austrian test-takers could easily identify, that might have enabled them to produce more conducive stimulated recall protocols than for the other two versions. Although the other two versions would have worked almost equally well in terms of stimulated recalls, it was decided to use this version for the main study. Second, in terms of coding category frequency for the stimulated recall protocols, no major differences between the two methods used in the pilot study were found. However, five of six participants clearly stated that Method 2 (replay of the eye-traces and mouse movements overlaid with the sound file) was more helpful to aid recall of their thought processes, as shown in the exemplary transcripts by two of the participants below. Only one participant found the answers as stimulus sufficient for recall. Based on these findings, it was decided to use Method 2 for the main study.

Researcher:	*Which of the two methods did you find more helpful for remembering what you were thinking: the one where I showed you the question and the answer you gave, or the one with the eye-movements?*
P01:	*The eye-movements.*
Researcher:	*Why?*
P01:	*Because I saw for myself what I was looking at, and somehow I find it easier [to remember what I was thinking] when I see that.*
Researcher:	*So which of the two methods did you find more helpful for remembering what you were thinking?*
P04:	*The method with the eye-tracking was more helpful for remembering my thoughts, because I was able to see where my focus was.*

Third, the inspection of the gaze plots revealed potential problems in the analyses of the eye-tracking data. The eye-tracking readings on the question stems were too low for some of the questions, and the readings on the four answer responses overlapped for a number of participants, due to inaccuracies commonly associated with eye-tracking (Holmquist et al., 2011). Based on these findings, the layout of the stimulus and the set-up of the eye-tracker (the seating position of participants and the tilt-angle of the eye-tracking screen) were changed slightly to mitigate these effects in the main study. Several test-runs were performed by the researchers before the main data collection to achieve satisfactory accuracy of eye-tracking readings.

4.2 Main study

4.2.1 Participants

Thirty participants, 16 female and 14 male, took part in the eye-tracking study. They were all German native speakers between 20 and 61 years of age (M=28.5) and had been learning English in class for 7 to 14 years (M=8.8). For most of the participants (N=21), their last English learning experience in class had been more than one year ago, and for one participant as long as 43 years ($M_{N=30}$=5.1). They had been living in an English-speaking country for between 0 to 20.5 years (M=2.5), with the majority (N=22) for not more than one year. Eight participants reported using English daily or almost daily, nine participants once or twice a week, three participants once or twice a month, and nine participants less than once a month (with one missing answer). In terms of level of education, one participant had finished compulsory school, two participants had finished vocational training school, 24 participants had graduated from upper secondary school, and three participants had obtained a university degree.

Of the 30 participants, 21 also performed stimulated recall protocols for each item. The remaining nine participants did not produce stimulated recalls due to time restraints. In addition, not all 21 participants were able to produce good quality stimulated recall data, so of the 21 stimulated recalls 16 were transcribed and included in the analysis. The decision on which of the stimulated recall protocols would be included in the analysis was based on the researchers' general impression on how well participants were able to recall their thoughts during the experiment. The 16 participants included in the stimulated recall analysis were between 20 and 61 years old (M=27.3). Nine participants were female and seven were male.

4.2.2 Materials

4.2.2.1 Aptis Listening Test

One Aptis Listening Test component consisting of 25 four-option multiple-choice items was used in the main study. As outline above, this component was chosen after the pilot study out of a total of three different versions, as it produced better quality stimulated recall data than two other versions. The test component included seven A1 items, seven A2 items, six B1 items, and five B2 items. The items were presented in order of difficulty, starting at A1 level, which is consistent with the operational Aptis Test.

Prior to data collection, the file format of the items had to be changed to be used for the Tobii Studio Pro eye-tracking software. All of the items were transformed into separate html files. The advantage of changing the file format was that the individual items' sound files could be intrinsically linked to each html file, so the start time of the sound files was standardised across all participants once they clicked the play button and did not have to be controlled externally by the researchers. This was important, as eye-tracking data is measured in seconds and different sound file start times would have muddied the output. Each item was programmed as a separate html file and imported into the Tobii Studio Pro software.

Due to the file format change, and to improve eye-tracking data quality, the layout of the items was changed slightly from the original Aptis layout. The question stem and the four answer options were aligned so that most of the eye-tracking screen was used, in order to minimise effects of potential eye-tracking inaccuracies. However, the general layout such as colouring, pictures, symbols, and type of font were the same as in the operational Aptis Test.

4.2.2.2 Measure of receptive English proficiency

Apart from the chosen Aptis Listening Test component used for the eye-tracking and stimulated recall investigation, participants also took a full set of the Aptis Reading, Listening, and Grammar and Vocabulary Tests as a measure of general receptive English language proficiency. It was hoped that the participants would be spread equally across the four CEFR levels targeted by the Aptis Test in terms of their general proficiency, and they were recruited to that end. In addition, the general receptive proficiency measure was used to answer the sub-questions 1.2b and 1.2c discussed in Section 5.2.1.2.

4.2.3 Procedure

4.2.3.1 Aptis Listening Test

The data was collected over two sessions. During the first session, all participants performed the eye-tracking experiment, and 21 out of 30 participants also performed stimulated recall protocols, as described above. Prior to the experiment participants first filled out a biodata questionnaire, followed by one of the researchers explaining in detail what was expected of them. Participants were familiarised with the task format by performing one example item on the Tobii TX300 eye-tracker.

They were then told that they had to answer 25 multiple-choice listening items and that they could listen to each recording twice. In line with the operational Aptis Test, participants could also choose to listen to the recording of an item only once, or move to the next item whenever they felt they knew the answer. Those who also performed stimulated recall protocols were told that they would be asked some questions after three to four items, and that their answers to the questions would be recorded on an audio recording device. Participants were reminded to treat the experiment like a normal language test. All instructions were given in German, which was the participants' native language.

Once all questions were answered, the participant's seating position was adjusted. Their left hand was positioned in a way that they were able to press the ESC key on the keyboard without needing to move their arm (the ESC key was used to move to the next item in case test-takers decided not to listen to the entire sound file, which is in line with the operational Aptis Test), and their right hand was placed on the mouse (this was inverted for left-handed participants). Once a comfortable position was obtained, an eye-tracking calibration was performed in Tobii Studio Pro, in order to find the best position for accurate eye-tracking readings. Following successful calibration, the participant was asked not to move their head anymore to maintain accurate eye-tracking readings throughout, and the first set of items was presented to the participant. Although the eye-tracker does allow for some natural head movement, too much movement could impact on the accuracy of the eye-tracking recording. Thus, although candidates were instructed to keep their head still, some natural movement was allowed. Participants did not feel obstructed by this in any way. The participant's eye-position was monitored throughout the test in the "Track Status" window of Tobii Studio. Participants who shifted outside the acceptable boundaries were instructed between items to move their position slightly so that eye-movements could be recorded accurately throughout. Items were presented in sets of three to four, after which participants could move freely again for the stimulated recalls. An eye-tracking calibration was completed before starting each new set.

For the stimulated recall sessions, the participant was asked to recall in detail how they had arrived at their answers for an item after each set of items. A recording of participants' eye-and mouse-movements overlaid with the sound file was replayed for each item to stimulate recall. The recording was stopped for recall at the point when the participant had chosen an answer. Another recall was initiated after the first play if a participant had decided to listen to the recording of an item a second time. If a participant had listened to an item twice, the recording was stopped again for recall at the point when they had changed their answer during the second play, if they had done so. The recording was also stopped if the participant showed obvious reactions to the recording, for example, when they laughed out loud or nodded, or when unexpected eye-movements occurred. As mentioned above, participants could move freely during the stimulated recalls and an eye-tracking calibration was performed each time before the next set of items was started. All stimulated recalls were conducted in German, but participants could also use English.

For each stimulated recall, the following standardised questions were asked in German:

> Before the first item in each set:
> *You are now going to see a recording of your eye-movements when you answered the last few items. While watching the video, try to remember what you were thinking while you answered the items. I will stop the video once or twice and will ask you some questions about it.*
>
> After the answer had been chosen during the first listening:
> *How did you arrive at this answer?*
> *Did you have any difficulties answering this item?*
> If yes: *Why was it difficult?*
> If no: *Why was it not difficult?*
>
> After the first play, if a participant had played the recording a second time:
> *Why did you decide to listen to the recording again?*
>
> After the answer had been changed during the second listening:
> *Why did you change the answer?*

When the participant showed particular reactions:

You laughed/nodded/agreed etc. Can you explain why? What were you thinking while you answered the item?

The procedure was repeated until all 25 items were completed (after a total of eight sets of three to four items each). In most cases, the sessions took between 1.5 and 2 hours, depending on how often participants decided to listen to the recording of an item more than once. Participants were given the opportunity to take a short break halfway through the experiment.

During data collection, a second researcher noted down the answers the participants gave to each of the items, and recorded the entire session on camera. The camera recording served as back-up in case the audio recording device failed, and it also aided subsequent transcription of stimulated recalls. For example, when participants referred to sections on the screen by pointing at it during recall, the video recording was used by the transcriber to identify what the participant was referring to.

4.2.3.2 Measure of receptive English proficiency

The full Aptis Reading, Listening, and Grammar and Vocabulary Test used as measure of general receptive English language proficiency was administered to each participant after the eye-tracking and stimulated recall protocol experiment in a separate session. It was ensured that a different listening section to the eye-tracking experiment was used. The use of a different test than Aptis as measure of general proficiency was not feasible in the time frame. Most participants performed the test within a few hours after the experiment, and a smaller number came in the following day. The test was administered on a computer using the original online testing tool by the British Council. The test administration guidelines provided by the British Council were followed. After successful completion of the test participants were paid 30 or 40 Euros for their time, depending on whether or not they completed a stimulated recall protocol during the eye-tracking experiment (the sessions with stimulated recall took longer).

4.2.4 Ethical consent

Prior to data collection, ethical approval for the project was obtained from the University of Innsbruck. A Certificate of Good Standing was granted by the Board for Ethical Issues at the University. Participants signed an information sheet outlining the study and filled out ethical consent forms before data collection. All participants agreed in writing to take part in the study.

4.2.5 Data analysis

4.2.5.1 Eye-tracking

Bearing in mind the limitations outlined in Section 2.1, most importantly the impossibility to disentangle reading and test-taking processes from listening processes in eye-tracking experiments involving complex listening stimuli, the quantitative eye-tracking analyses presented in this study were intended to be exploratory. The primary explanatory variable chosen for the analysis was visit duration on response options, i.e. the amount of time participants spent looking at each of the four response options. This was because during the eye-tracking data collection, it was evident that the participants primarily focused on the response options, so the majority of cognitive processing in relation to eye-movements would have occurred then.

Prior to the quantitative eye-tracking data analysis, the following underlying hypothesis was formulated: *As the CEFR-level of the items in the Aptis Listening Test increases, the time test-takers spend looking at each of the four response options increases as well.*

This hypothesis was based on the fact that lower-level items target a) lower-level listening processes and b) more local type of information, according to the Aptis test specifications, than higher-level items. In other words, it was hypothesised that test-takers would need to spend less time looking at the four response options for lower level items, as lexical search processes of lexical information for A1 items would necessitate lower visit durations on the four responses in order to find the correct answer, than discourse construction processes of meaning at the discourse level for B2 items.

In addition, participants would need to employ a larger amount of cognitive processing overall for higher-level items, as higher-level listening processing by definition is based on lower level processing (Field, 2013).

Thus, if this hypothesis were confirmed, it would be indirect evidence of the cognitive validity of the Aptis Listening Test. It needs to be stressed at this point, however, that the time test-takers spend on looking at the response options of an item does not indicate the amount of *listening* processing *per se*. As Winke and Lim (2014) rightly point out, the construct of listening and reading cannot be disentangled with eye-tracking measurements in listening test experiments. Rather, it is hypothesised that eye-tracking metrics can be indicative about the *overall* amount of cognitive processing, which subsumes listening, reading and test-taking processing. Still, by correlating test-takers' listening and reading scores on the Aptis test package used as measure of general proficiency with the eye-tracking metrics captured in the study, it was hoped to gain some insights into the relative importance of the two constructs. Furthermore, during completion of the items, processing will also have been occurring while participants were not focusing on a particular response option. However, given that this cannot be measured with eye-tracking, it will have to be overlooked in this part of the research study, though this is accepted as a limitation.

Based on this discussion and the research questions outlined in Section 3, the following sub-questions were formulated:

RQ1.2a *To what extent do test-takers spend more time looking at the response options of higher-level items as compared to lower-level items in the Aptis Listening Test?*

RQ1.2b *To what extent does test-takers' listening ability have an impact on the time they spend looking at the response options in the Aptis Listening Test?*

RQ1.2c *To what extent does test-takers' reading ability have an impact on the time they spend looking at the response options in the Aptis Listening Test?*

To compare visit durations on response options with the CEFR level of the items and participants' scores on the Aptis Reading and Listening Tests used as measure of general proficiency, a number of variables needed to be controlled for in the analysis, as outlined in Table 4.

Variable name	Technical description	Why included?
Visit duration	The visit duration, measured in seconds, for a particular individual on a particular item on a particular response. **(dependent variable)**	It is assumed that this is a measure of the overall amount of cognitive processing undertaken.
Listen times	The number of times the participant listened to the text of a particular item (once or twice) **(control variable)**	In the Aptis Listening Test participants can choose whether to listen to the text of an item once only or twice. Therefore, the number of listening times needed to be controlled for in the analysis before comparing visit durations between items of different CEFR levels.
Response order	The order in which the responses were presented on the screen **(control variable)**	During the exploratory analysis of the eye-tracking data, a strong tendency for participants to focus more on the response options that were presented higher up on the screen was detected. This tendency needed to be controlled for when comparing the amount of time taken to look at items of different CEFR levels, as the location of the correct response was not equally balanced across all levels.
Response chosen	A binary indicator of whether the particular response option was chosen by the participant **(control variable)**	This variable controlled for the fact that participants put additional focus on the option chosen due to: 1) the processing which occurs when matching the text to the representation of the chosen answer; and 2) the need to execute fine motor control with visual feedback to click the mouse in the correct location.

Table 4: Description of dependent and control variables

The following variables were relevant to answer the research questions.

Variable name	Technical description	Why included?
CEFR item	The British Council assigned CEFR level of the item	As the main explanatory variable, it was hypothesised that visit duration would increase as the CEFR level of the items increases, which would indicate a need for more cognitive processing on more difficult items.
Listening score	The raw Aptis Listening Test score of the participant used as a measure of general English proficiency	This variable was included to investigate how listening ability impacts the amount of time spent looking at the textual information in the responses.
Reading score	The raw Aptis Reading Test score of the participant used as a measure of general English proficiency	This variable was included to investigate how reading ability impacts the amount of time spent looking at the textual information in the responses.
Eye-tracking score	The raw score for the specific Aptis Listening Test items responded to during the data collection	This variable was included to investigate how the overall score on the particular set of items used in the study predicted the amount of time focusing on the textual information in the responses.
Participant	A factor indicating which participant the visit duration came from	This variable was included to model the proclivity of a particular person to focus on the responses.
Item	A factor indicating which item the visit duration came from	This variable was included to model the proclivity of a particular item to elicit focus on the responses.

Table 5: Description of exploratory variables

In order to analyse the data on the total visit durations for each response, by each person and on each item, a regression model needed to be implemented. Regression models have the advantage that a number of variables can be modelled jointly, while simultaneously controlling for different variables without the need for voluminous amounts of averaging across variables. As outlined in Table 4, the specific variables that needed to be controlled for by the model presented here were the number of times a participant listened to the text (participants had the option of listening once or twice, and if they listened twice they had more opportunities to look at the responses), the ordering effect of the items (it was clearly seen that participants tended to focus more on items higher up the page), and the particular response chosen by the participant (participants focused more on the chosen response as they had to carefully manoeuvre the mouse to a small area to choose their answer). If these factors were not accounted for, there would be a risk of confounding variables and making incorrect inferences.

A mixed effects linear regression model (Gelman & Hill, 2006) with random intercepts was chosen as the best option for data analysis. Mixed effects linear regression models have been successfully employed for research in linguistics and have been suggested as a useful method by a number of researchers (e.g. Baayen, Davidson & Bates, 2008; Winter, 2013). A mixed effects model divides the explanatory variables into two kinds, *fixed effects* and *random effects*. In simple terms, *fixed effects* are parameters of the model that are of interest to the investigation, whereas *random effects* are factors that are incidental and, randomly, come from a large population. In the model presented in this report, the fixed effects were: response order, listen times, response chosen, CEFR item, listening score, reading score, and eye-tracking score. These variables are informative about how different attributes of the specific response option, the specific item or the specific person go about affecting the amount of cognitive processing on a given response option. The variables participant and item were designated as random effects. The visit durations made by individual participants or elicited by individual items were not relevant to answer the research questions, as the specific participants or specific items in the experiment could effectively be replaced and information on cognitive processing in relation to the fixed effects would still be procured. However, clearly, there will be a proclivity for specific participants and items to make and elicit differing visit durations. In other words, the visit durations measured on a specific participant or item will be correlated. Mixed effects models take this correlation into account in a way that allows us to generalise to the population of participants and items.

If the correlation between the visit durations of participants and items was ignored, and a regression model that does not take these random effects into account was fitted, researchers would run a risk of drawing spurious conclusions about the statistical significance of the fixed effects in the model (Crawley, 2007).

In order to undertake a linear mixed effects analysis of the relationship between the various factors relating to participants, items and specific responses to the total visit duration on the specific responses, package lme4 (Bates, Maechler, Bolker, & Walker, 2015) for R (R Core Team, 2014) was used. The random effects, participant and item, were characterised by a random intercept. Visual inspection of the residual plots did not reveal any serious violations of the assumption of normality and homoscedasticity. Only one outlier, of three thousand data points, was removed from the model as the residual was suspect. The p-values for the fixed effects were obtained via Satterthwaite approximation using the package lmerTest (Kuznetsova, Brockhoff & Bojesen Christensen, 2016).

4.2.5.2 Stimulated recall

As outlined above, 16 candidates were chosen to produce stimulated recall protocols based on a replay of the videos of their eye movements overlaid with the sound file of the items. The protocols were transcribed based on the audio recordings by a research assistant. The verbal protocols were kept in the original languages chosen by the participants, i.e. German or English or, in most cases, a mixture of both. Transcripts were not translated. The video recordings were only used when there was a need for clarification due to technical problems with the sound file or when it was evident from the sound file that participants had pointed to something in particular on the computer screen.

The transcripts were then coded in the qualitative data analysis software Atlas.ti v7, both for the cognitive processes according to the model by Field (2013), and the information used as targeted by the Aptis test specifications (O'Sullivan & Dunlea, 2015). One researcher coded the cognitive processes, and another researcher coded the information used.

Field's (2013) model of cognitive processing in listening provided the basis of the coding framework for the cognitive processes used by test-takers. The Aptis test developers specifically refer to this model as the basis for their test construction. Also, since they follow Weir's (2005) socio-cognitive test validation framework, Field's cognitive processing model appeared to be an appropriate theoretical basis. One extra code was added during the coding for the few instances when candidates admitted to "pure guessing". This approach is in line with Brunfaut and McCray's (2015) research.

In addition, the "information used" as specified by the Aptis test developers were coded to check whether the test items successfully operationalise the targets per level (O'Sullivan & Dunlea, 2015). The codes are listed in Table 6.

1 Cognitive processes	2 Type of information used	3 Free codes
1.1 Input decoding	2.1 Lexical recognition	3.1 Pure guessing
1.2 Lexical search	2.2 Factual information	
1.3 Parsing	2.3 Interpretive meaning at utterance level	
1.4 Meaning construction	2.4 Meaning at the discourse level	
1.5 Discourse construction		

Table 6: Coding scheme for coding transcribed stimulated recall data

To facilitate and standardise the coding, a fuller scheme was developed by the research team based on short theoretical definitions of each code from the literature and illustrative examples from the pilot data. Furthermore, the transcript for each item was coded for the correctness of the final answer of the respective candidate on the test. Items were therefore coded "correct" or "incorrect", so that only correctly answered items could be easily extracted for further analyses. This is in line with Brunfaut and McCray, who report that a "distinction was […] made between codings associated with correctly answered items and those associated with incorrectly answered items" (Brunfaut & McCray, 2015). Following this approach, the focus of all sub-analyses for validation purposes was mainly on processes associated with correctly completed items. This was done because it seemed appropriate for this kind of validation research to only investigate the processes used by candidates who managed to respond to items correctly. Also, it appeared to make for relevant comparisons between successful and unsuccessful candidates in terms of their cognitive processes employed.

In contrast to Brunfaut and McCray's (2015) study on the Aptis Reading Test, the total number of occurrences of each of the coding categories (Table 6) of all test-takers was not calculated. Instead, it was only calculated whether or not a particular process had been evidenced in the verbal report per candidate and item. It was disregarded whether any process was referred to multiple times in one item transcript. This way, a more standardised comparison between the items and candidates was enabled, as transcripts usually differ considerably in length.

The data was probed further to answer the subordinate research questions and investigate differences in cognitive processing and in information used of items targeting different CEFR levels (RQ 1.2 and RQ 2.2). For this, each item was given a code based on the stipulated CEFR level of the item by the test design team.

5. RESULTS

5.1 Descriptive statistics

5.1.1 Aptis Listening Test and measure of receptive English proficiency

Table 7 shows the results of the participants on the different test packages used in the investigation: the Aptis Listening Test version used for the eye-tracking and stimulated recall study, as well as the Aptis Listening, Reading and Grammar and Vocabulary Tests used as measure of the participants' receptive proficiency. As can be seen, the participants performed well on all of the tests used in the investigation.

	Maximum score	Minimum	Maximum	Mean	Standard deviation
Listening (eye-tracking and stimulated recall)	25	15	25	21.63	2.58
Listening	50	28	48	42.07	5.50
Reading	50	28	50	45.53	5.75
Grammar and Vocabulary	50	22	47	36.83	6.77

Table 7: Descriptive statistics of the different test packages

In terms of CEFR level, as mapped by the British Council for the computer-delivered Aptis Listening and Reading Tests used as measure of receptive proficiency, most of the participants were placed above B2 level, as shown in Table 8. Although it was aimed to recruit participants in the A2 to C range, these results show that most participants were proficient users of English in terms of their receptive abilities.

	A1	A2	B1	B2	C
Listening	0	0	2	2	26
Reading	0	0	2	4	24

Table 8: CEFR levels of participants

5.1.2 Listening once vs. listening twice in the Aptis Listening Test

As outlined in Section 4.2.3.1, in the Aptis Listening Test candidates can choose whether to listen to the recording of an item once or twice. Unfortunately, for the computer-delivered version of the Aptis Test used as measure of receptive English proficiency, this information is not reported by the British Council. However, we were able to capture whether candidates listened to the recordings of the individual items once or twice in the Aptis Listening Test used for the eye-tracking and stimulated recall experiment. Table 9 shows the total number of times candidates listened to items once or twice in relation to the items' CEFR level.

It can be seen that there is a clear relationship between the CEFR level of an item and the sum total of times the recording was listened to once or twice. As the CEFR level increased, the observed proportion of times the items were played twice increased as well. The correlation, as measured by Spearman's Rank Order Correlation, between the item CEFR level and the number of times an item of that CEFR level required a second listen, across all 30 participants, was ρ =0.55 (p=0.00**).

	A1 (7 items)	A2 (7 items)	B1 (6 items)*	B2 (5 items)	Total (25 items)
Listening once	201	185	113	50	549
Listening twice	9	25	66	100	200

*Participant 26 accidentally skipped item 20, so the total does not add up to 180

Table 9: CEFR level by number of times listened to the recording for the Aptis Listening Test

5.2 Eye-tracking

5.2.1 Linear mixed model

As outlined in Section 4.2.5.1 above, due to the impossibility of disentangling listening processes from reading and test-taking processes with eye-tracking measurements in complex listening tests, the quantitative eye-tracking analysis presented here was exploratory. It was hypothesised that with increasing item level, the time test-takers spend looking at the four response options would also increase. In order to analyse the visit duration measurements on the four response options in relation to the items' CEFR level, a linear mixed model needed to be implemented. The results of the linear mixed model are presented in this section.

The dependent variable in the model – visit duration – was highly negatively skewed (see Figure 1 below). Various options were available for modelling the data. It would have been possible to analyse the data using a Gamma distribution, and this was found to be a good fit. However, for reasons of interpretability, it was chosen to perform a natural log transformation on the raw value for visit duration (see Figure 2 below). The advantage of a log transformation is that the coefficients of the regression model can be interpreted in terms of percentage increase or decrease in visit duration under the different conditions. While there was still skew in the transformed data, it was found suitable for regression analysis as one of the main assumptions of regression is a normal distribution of the residuals from the model.

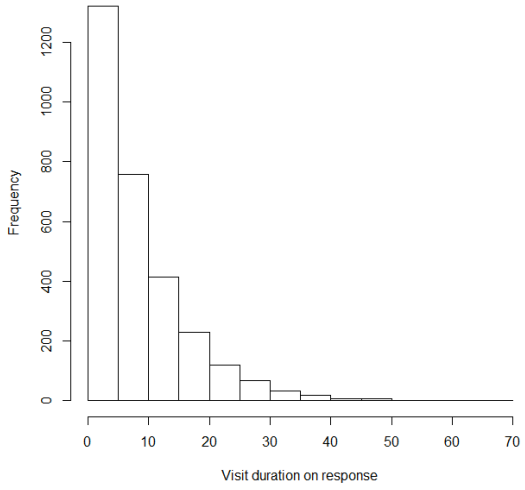

Figure 1: Histogram of visit duration on response options

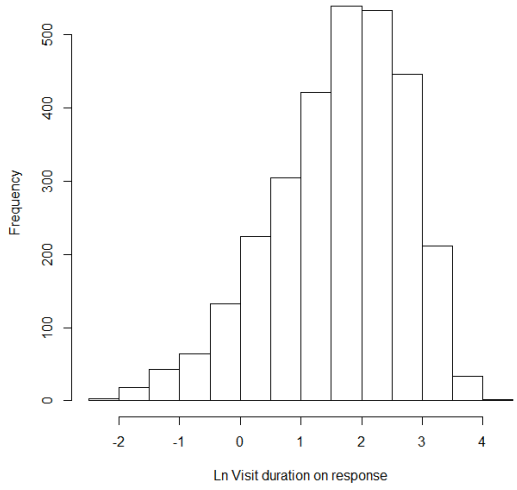

Figure 2: Transformed distribution of visit durations on response options

Table 10 shows the results of the linear mixture model. In terms of Akaike Information Criterion (AIC) (Akaike, 1974), a commonly used measure of model parsimony, the inclusion of both random effects, participants and items, was more parsimonious than the inclusion of just one or none. In other words, both random effects should be included in the model. It can be seen that the participant and item random variables do explain some of the total variance, but that the majority remains as unexplained (residual) variance. The squared correlation between the observed and fitted values for the model is 0.68 (bottom of Table 10), showing that the model explains a substantial amount of the variance in the visit durations. For the fixed effects, the raw β estimate (a percentage change effect, described below), its standard error, the approximate degrees of freedom and the associated p-value are reported.

Random Effects	**Variance**	**Std. dev.**		
Participants	0.06	0.25		
Items	0.08	0.28		
Residual	0.39	0.62		
Fixed Effects	**β estimate**	**Std. error**	**Approx. df**	**p-value**
(intercept)	2.23 (9.31s)	0.44	30	0.00***
response order 2	-0.38 (-30%)	0.03	2913	0.00***
response order 3	-0.82 (-56%)	0.03	2914	0.00***
response order 4	-1.39 (-75%)	0.03	2914	0.00***
listening twice	0.09 (+9%)	0.03	2941	0.01**
chosen response	0.79 (+120%)	0.03	2914	0.00***
CEFR A2	0.34 (+40%)	0.15	21	0.04*
CEFR B1	0.90 (+145%)	0.16	21	0.00***
CEFR B2	1.22 (+240%)	0.17	22	0.00***
listening score	-0.03 (-3%)	0.01	26	0.06
reading score	0.03 (+3%)	0.01	26	0.04*
eye-tracking score	-0.04 (-3%)	0.03	26	0.29

Squared correlation between observed and fitted (pseudo r^2) = 0.68

Table 10: Results of the linear mixed model

Under normal circumstances, i.e. when the dependent variable has not been log transformed, there are differing interpretations of categorical and continuous explanatory variables based on the β estimates in Table 10. Specifically, for the continuous variables, a 1-unit increase (or decrease) in that variable represents a "β estimate" sized increase (or decrease) in the dependent variable, and for the categorical variables the β estimate represents the amount that should be added (or subtracted) from the intercept when the data point comes from that specific category.

As mentioned above, the dependent variable in this model was log transformed and therefore a slightly different interpretation of the β estimate is required. Specifically, the value of interest is the exp(β Estimate), where "exp" is the exponential function – the inverse of the natural logarithm. The exp(β Estimate) can be translated into a percentage increase or decrease for either a 1-unit increase in the explanatory variable or the effect of belonging to that category, for continuous or categorical variables, respectively. Table 10 gives the percentage increase or decrease in the expected visit duration in parentheses after the β estimate. The exponential of the intercept (9.31 seconds) is the fitted total visit duration on a particular item's response that was first on the page, where the text was listened to once, that was not the response chosen by the particular participant, that was CEFR level A1, by a participant with listening, reading and eye-tracking test scores of 0. In the following, the results on the different fixed effects included in the model as displayed in Table 10 will be described.

As can be seen in Table 10, the response order shows an unambiguous and highly statistically significant pattern: the lower the response on the page, the lower the participants' visit duration. The participants' total visit duration was 30% lower for the second response option than the first, 56% lower for the third response option than the first, and 75% lower for the fourth response option than the first. If this variable had not been controlled for in the model, the risk of concluding that items at different CEFR levels elicited different amounts of cognitive processing may have been confounded with the correct response locations for the items of a particular CEFR level. In other words, if the correct response location for all A1 items had been location 1 (top of the page) and for all A2 items location 4 (bottom of the page), the model might have shown that A1 responses elicited longer visit durations than A2 responses, as the effect of the correct response being at the top of the page would have generated more visits overall for A1 items.

When all other variables are controlled for in the model, the effect of listening to the text twice increased the visit durations on response options by 9%, statistically significantly. This is surprisingly low. Given Table 9 above, it is likely the case that this low coefficient is explained by the correlation with CEFR level. In other words, the fact that higher CEFR level items generate more instances in which the text is listened to twice means that a large proportion of the variance in whether the text is listened to twice is accounted for by the CEFR level. However, this variable still needs to be included as a predictor to control for the fact that whether a participant listened to a text once or twice was not under experimental control.

The particular response a participant chose increased the total visit duration on that response by 120%, as shown in Table 10. This seems logical, as the participant had to perform the fine motor task of clicking the mouse in the appropriate place on the screen to answer the item correctly, and likely performed additional matching processing on that particular response before confirming their selection. Again, it is important that this variable is controlled for as it could not be manipulated experimentally, potentially leading to confounding with response order, for reasons discussed above, and thus with CEFR level.

5.2.1.1 Relationship between visit duration and item level (RQ1.2a)

The CEFR levels of the items, as the main explanatory variables in Table 10, show a clear pattern. With all other variables being controlled for, A2 items elicited 40% longer visit durations than A1 items, B1 items elicited 145% longer visit durations than A1 items, and B2 items elicited 240% longer visit durations than A1 items. This is evidence to suggest that higher-level items elicit more processing on the item responses than lower-level items. However, it needs to be stressed again that it is assumed that visit duration on the response options is not indicative about the amount of listening processing, but rather on the overall amount of processing, including listening, reading, and test-taking processing.

5.2.1.2 Relationship between visit duration and listening and reading ability (RQ 1.2b and RQ1.2c)

The regression coefficients which relate visit durations to the measures of listening and reading ability displayed in Table 10 (fixed effects at the bottom of the table: listening score, reading score, and eye-tracking score) provide interesting results from which tentative conclusions can be drawn. An increasing listening score, and an increasing score on the stimulus items (eye-tracking score), suggest a 3% reduction in visit duration on the response options per unit score, although not statistically significantly. Conversely, a 1-unit increase in score on the reading items suggests a statistically significant increase in visit duration on the responses. A possible conclusion that could be drawn from this pattern is that increasing listening ability might mitigate the need to focus on responses, while increasing reading ability might actually increase the utility of the written response in the processing around the construction of an answer to the item. However, it could also simply mean that better readers read the answer options more often, without there being an effect on how they construct an answer to the item. Also, the amount of information available to provide evidence for this assertion is relatively low (26 df, see Table 10), and the actual scoring data showed strong ceiling effects. This finding should be interpreted with these factors in mind.

5.3 Stimulated recall

To further investigate the four main research questions, stimulated recall data was gathered. A complete set of an Aptis Listening Test with 25 items was administered to 16 candidates, and their recorded eye-movements overlaid with the sound file of the items were used as stimuli immediately after taking sets of three to four items. These verbal protocols were then coded according to the taxonomies suggested by Field for cognitive processes and the Aptis test specifications for information used (see Section 4.2.5.2). In total, the 16 candidates answered 88.5% of the items correctly (354 of 400 cases), with only 11.5% incorrect answers (46 cases).

In the following section, the findings will be presented according to the four research questions. For each research question, the overall counts will only be presented for the processes and information used for correctly answered items. As indicated in the methodology section, the tables do not illustrate the overall number of instances where any one process or information used was verbalised, but are displayed in clustered form, i.e. whether or not the particular process or information used was identified in the transcripts at least once per candidate per item.

Percentages have then been calculated given the maximum number of possible cases. However, test-takers often indicated more than one type of cognitive process employed to arrive at an answer. They also sometimes relied on more than one type of information used. For this reason, percentages across the columns do not add up to 100%. Unfortunately, the quantitative data cannot be illustrated with quotes from the stimulated recalls in this report because the listening items used in the study are live test items, the content of which cannot be disclosed.

5.3.1 Cognitive processes used to answer items on the Aptis Listening Test (RQ1.1)

Table 11 illustrates that the complete Aptis Listening Test appears to tap into the whole range of targeted cognitive processes. The results indicate that for correctly answering the Aptis Listening Test, candidates employed lower-level processes such as input decoding, lexical search and parsing, as well as higher-level processes such as meaning and discourse construction. Overall, the most used processes were lexical search and meaning construction. There was only marginal evidence for input decoding processes employed in the stimulated recall data. This, however, may be in line with Brunfaut and McCray's (2015) findings for reading, where the authors found no evidence for the lowest level process of word recognition. Similar to Brunfaut and McCray's results on word recognition, the automated nature of input decoding in listening, particularly at the relatively high proficiency level of the study participants, may be an explanation for this finding. It should also be noted that "absence of evidence is not evidence of absence" (Godfroid & Spino, 2015, p. 896). Thus, it could be argued that the Aptis Listening Test requires candidates to employ the entire spectrum of processes specified in the model by Field (2013). In a limited number of instances, test-takers indicated that they had determined the answer based on guessing. While these instances were coded with an additional coding category, they were so rare that they were excluded from the data analyses here.

	1 Input decoding	2 Lexical search	3 Parsing	4 Meaning construction	5 Discourse construction
Item 1	0	88	50	13	0
Item 2	0	69	44	44	0
Item 3	0	81	38	13	6
Item 4	0	81	44	31	6
Item 5	0	88	50	13	6
Item 6	0	94	25	38	13
Item 7	0	88	50	13	13
Item 8	0	50	50	69	50
Item 9	0	38	50	69	19
Item 10	0	19	25	88	56
Item 11	6	25	31	88	50
Item 12	0	6	19	94	63
Item 13	6	19	13	94	63
Item 14	0	19	6	100	75
Item 15	6	25	25	69	31
Item 16	0	31	19	100	25
Item 17	6	13	6	44	13
Item 18	13	13	6	81	81
Item 19	6	25	31	81	38
Item 20	0	19	25	69	44
Item 21	0	0	0	38	31
Item 22	6	6	38	81	25
Item 23	0	0	0	75	69
Item 24	6	25	38	75	31
Item 25	0	0	6	63	50

Table 11: Cognitive processes employed to answer Aptis Listening Test items (correctly answered items only, in percent of cases)

5.3.2 Cognitive processes at the different item levels of the Aptis Listening Test (RQ 1.2)

To answer RQ 1.2, the results were grouped by item target level. This was done to identify whether the items at different levels trigger different cognitive processes in order to elicit successful answers. While it may be expected that the different task types used in the Aptis Reading Test would elicit different processes at different levels, as found by Brunfaut and McCray (2015), it seemed important to explore whether the Aptis Listening Test, employing multiple-choice items only, would also elicit different processes at different CEFR item levels.

For CEFR items levels A1 and A2, the analysis included data on seven items each, totalling 112 answers given by the 16 participants for each level. For B1, the data set consisted of six items (96 answers) and for B2 it comprised five items, resulting in 80 answers.

Participants gave 112 correct answers on the A1 items (100%), 110 correct answers on the A2 items (98%), 76 correct answers on the B1 items (79%) and 56 correct answers on the B2 items (70%). Table 12 shows the participants' cognitive processing per target CEFR level of the tasks, as evidenced in the stimulated recall data. Data for all answers is presented in Table 12 and only for correct answers in Table 13.

	1 Input decoding	2 Lexical search	3 Parsing	4 Meaning construction	5 Discourse construction
A1 (112)	0.0	83.9	42.9	23.2	6.3
A2 (112)	3.6	25.9	28.6	86.6	53.6
B1 (96)	5.2	28.1	25.0	93.8	53.1
B2 (80)	2.5	15.0	28.8	96.3	60.0

Table 12: Cognitive processes employed per target CEFR level (overall, in percent of cases)

	1 Input decoding	2 Lexical search	3 Parsing	4 Meaning construction	5 Discourse construction
A1 (112)	0.0	83.9	42.9	23.2	6.3
A2 (110)	1.8	25.5	28.2	87.3	54.5
B1 (76)	6.6	26.3	23.7	93.4	48.7
B2 (56)	3.6	8.9	23.2	94.6	58.9

Table 13: Cognitive processes employed per target CEFR level (correct items only, in percent of cases)

It seems surprising that the lower-level process of input decoding is not evidenced at the lowest proficiency level items, but at the three higher levels. However, the evidence for this processing is minimal even at these levels, which is likely to be due to the nature of the process and the research methodology not being suitable to tap into or make visible this kind of automated processing.

A1 items

For items at A1 level, candidates mostly relied on lexical search processes. To arrive at the correct answer on these items, they reported on activating lexical search in 83.9% of the cases. However, candidates also evidenced employing processes such as parsing (42.9%), as well as the higher-level processes of meaning construction (23.2%) and discourse construction (6.3%).

This may be somewhat surprising, as A1 items do not target higher-level processes. A1 items aim primarily at activating lower-level processes such as lexical search or input decoding, the latter of which is not evidenced in the stimulated recall data at all. While this lack of evidence for input decoding might be for the reasons outlined above, the evidence for the instances of higher-level processes could be explained by the relatively high proficiency of the candidates in relation to these items. For example, a proficient test-taker would evidence discourse construction when re-narrating the entire sound file in the stimulated recall. This evidence does not, however, imply that higher-level processing would be necessary for successful completion of A1 CEFR-level items. This finding is in line with Brunfaut and McCray's (2015) results on the Aptis Reading Test, where the authors report that candidates often relied on more than one type of cognitive processing to arrive at the correct answer for an item. For most cases of A1 items in our study, this was a combination of lexical search with either parsing or meaning construction.

A2 items

To answer A2 items correctly, participants mostly employed meaning construction processes (87.3%). They often did this in combination with discourse construction (54.5%), parsing (28.2%), and lexical search processes (25.5%). This finding is in line with Field's model and the Aptis Listening Test specifications, as items at this CEFR level are intended to mainly target meaning construction and discourse construction.

B1 items

The dataset for correctly answered B1 items looks very similar to that of the A2 items. This again seems reassuring, as similar processing is intended to be elicited by the Aptis test designers between these two item levels. Most candidates evidenced using meaning construction processes (93.4%) for arriving at the correct answer. Frequent instances of discourse construction (48.7%), lexical search (26.3%) and parsing (23.7%) could also be observed in the data. Surprisingly, for items at this level, some evidence of input decoding was also found in the dataset (6.6%). However, these instances were few in number and always occurred in conjunction with some other, higher-level, processing type. It is clear from the data that the lower-level process of input decoding alone is insufficient to answer items at this level correctly.

B2 items

For the majority of correctly answered B2 items, candidates reported using meaning construction processes (94.6%). They mostly used these in conjunction with discourse construction (58.9%) and to a smaller degree with parsing (23.3%). Lower-level processes such as input decoding (3.6%) and lexical search (8.9%) were evidenced only rarely in the stimulated recall data. The items at B2 level, despite being of the same item type as items of the lower levels (multiple-choice), appear to successfully elicit a larger amount of higher-level processes.

Item clusters across the four CEFR levels

When grouped by CEFR level, it emerges from the data that each cluster of items at any level elicits a range of cognitive processes related to listening. In terms of Field's (2013) taxonomy, both lower- and higher-level processes are employed by candidates to arrive at the correct answers. The ratio of processes evidenced in the stimulated recalls thereby differs as outlined in the Aptis test specifications. On the face of it, the findings corroborate that correct answers in A1 items are primarily associated with lexical search processes, while correctly answering A2 and B1 items relies mostly on meaning construction and discourse construction processes. At B2 level, items also appear to elicit mainly meaning construction and discourse construction processes, the proportion of the latter being higher than for lower-level items.

However, to investigate whether these differences in ratios were significant, we conducted a series of tests between all pairs of CEFR levels for cognitive processing using an exact pairwise Fisher test. An exact test was used as some of the cells have values of 0, meaning tests using an asymptotic approximation would not be valid. The procedure used is "pairwise.fisher.test" from the R package "FMSB" (Nakazawa, 2015). The Holm (1979) method of correction of type I error rate for multiple comparisons was used. A table of p-values for the pairwise comparisons is presented below for the cognitive processes found in protocols for correctly answered items.

	1 Input decoding	2 Lexical search	3 Parsing	4 Meaning construction	5 Discourse construction
A1 – A2	0.73	0.00	0.10	0.00	0.00
A1 – B1	0.06	0.00	0.05	0.00	0.00
A1 – B2	0.55	0.00	0.08	0.00	0.00
A2 – B1	0.55	1.00	1.00	0.54	0.92
A2 – B2	1.00	0.039	1.00	0.54	0.92
B1 – B2	1.00	0.039	1.00	1.00	0.87

Table 14: p-Values of pairwise comparisons of cognitive process ratios across CEFR levels (correct items only)

As can be seen in Table 14, only the CEFR level A1 items appear to be significantly different in terms of the cognitive processes employed from the other levels. For levels A2–B2 there is little evidence of statistically significant differences in the process ratios. The only exception is the process type, lexical search, which was employed in this dataset significantly less at B2 level than at A1, A2 or B1 level. This lack of significant difference may, in part, be attributable to the uniform item type that is being used in the Aptis Listening Test across all levels. As Brunfaut and McCray (2015) pointed out, "any differences [in amount and type of processes being used] may also be due to, or influenced by, task type, and it is indeed likely that the task formats partially explain the cognitive processing differences between the CEFR groups of items" (p.42). Therefore, the findings may suggest that different task types might be more suitable to activate the desired type of processing at the various levels.

5.3.3 Types of information used to answer items on the Aptis Listening Test (RQ2.1)

To answer Aptis Listening Test items correctly, candidates used the whole range of types of information outlined in the Aptis test specifications, as illustrated in Table 15. Compared to the types of cognitive processes the picture appears clearer, with any one item mostly requiring only one or two types of information. The results suggest that the test successfully incorporates items that require candidates to mainly rely on lexical recognition, understand factual information, arrive at the answer by interpreting meaning at the utterance level and, albeit in rather few cases, understand meaning at the discourse level.

Overall, the type of information evidenced most, by a large margin, in the stimulated recall data was "understanding factual information". It appears that most items in the Listening Test require some sort of comprehension of this type of information to arrive at the correct answer. In comparison to the cognitive processes, there also appears to emerge a clearer pattern of different information types being associated with different CEFR levels.

To explore this further, the items were clustered by target CEFR level in the following analysis to answer research question 2.2.

	1 Lexical recognition	2 Factual information	3 Interpretative meaning at the utterance level	4 Meaning at the discourse level
Item 1	88	13	0	0
Item 2	69	50	0	0
Item 3	89	19	0	0
Item 4	69	38	0	0
Item 5	81	31	0	0
Item 6	81	38	0	0
Item 7	69	44	0	0
Item 8	0	94	0	0
Item 9	13	81	0	0
Item 10	0	94	0	0
Item 11	0	100	0	0
Item 12	0	100	0	0
Item 13	13	100	6	0
Item 14	0	100	0	0
Item 15	0	63	13	0
Item 16	0	88	13	13
Item 17	0	38	19	0
Item 18	0	81	0	6
Item 19	0	75	13	0
Item 20	0	75	19	0
Item 21	0	13	13	25
Item 22	0	44	56	0
Item 23	0	19	38	31
Item 24	0	0	56	25
Item 25	0	6	25	38

Table 15: Information used to answer Aptis Listening Test items (correctly answered items only, in percent)

5.3.4 Types of information at the different item levels of the Aptis Listening Test (RQ2.2)

Table 16 and Table 17 present the results of the grouped analysis of all items and correctly answered items respectively. However, the findings will be described in detail for Table 17 only.

	1 Lexical recognition	2 Factual information	3 Interpretative meaning at the utterance level	4 Meaning at the discourse level
A1 (112)	77.7	33.0	0.0	0.0
A2 (112)	3.6	96.4	0.9	0.0
B1 (96)	0.0	83.3	18.8	3.1
B2 (80)	0.0	27.5	50.0	30.0

Table 16: Information used per target CEFR level (overall, in percent of cases)

	1 Lexical recognition	2 Factual information	3 Interpretative meaning at the utterance level	4 Meaning at the discourse level
A1 (112)	77.7	33.0	0.0	0.0
A2 (110)	3.6	97.3	0.9	0.0
B1 (76)	0.0	88.2	15.8	3.9
B2 (56)	0.0	23.2	53.6	33.9

Table 17: Information used per target CEFR level (correct answers only, in percent of cases)

A1 items

The stimulated recall protocols evidenced that in answering items at A1 level candidates mostly relied on lexical recognition. In 77.7% of the cases, this type of information was used to arrive at the correct answer. In only 33.3% of the cases, the items appeared to require understanding of factual information. No test-taker reported using interpretative meaning at the utterance level or having to understand meaning at the discourse level for the seven items at this level. This is largely in line with the Aptis Listening Test specifications, although it could be argued that the items evidencing the use of factual information may not be targeting exactly what is intended. It does seem, however, that lexical recognition is sufficient in the majority of cases to successfully complete these A1 items.

A2 items

For the A2 items, the results are fairly similar to the A1 items, with lexical recognition and factual information almost exclusively being used by test-takers to answer items correctly. However, a progression can be observed in that the ratio is now inverted for this group of items. Almost all A2 items required understanding of factual information (97.3%). Only minimally this was done in combination with lexical recognition (3.6%) and interpreting meaning at the utterance level (0.9%). Candidates' stimulated recalls therefore confirm the intended targets of this group of items, as the specifications claim that factual information is the main type of information targeted at this CEFR level.

B1 items

To answer the B1 items correctly, candidates heavily relied on factual information (88.2%). There was also evidence in the stimulated recall data at this item level for some use of interpretative meaning at the utterance level (15.8%) and a few instances of meaning at the discourse level (3.9%). This again indicates a fairly clear progression from the A2 level items, while still being in accordance with the specifications, which state that items at B1 level should predominantly target factual information as type of information required to answer items correctly.

B2 items

Items at B2 level seem to mainly target different types of information than items at the other levels. While candidates still report some use of factual information (23.2%), the primary type of information required by this group of items was interpretative meaning at the utterance level (53.6%). Also, meaning at the discourse level (33.9%) appears to be tested at this CEFR level, as evidenced in the stimulated recalls. Again, this finding can be argued to be in line with the Aptis Listening Test specifications.

Item clusters across the four CEFR levels

While the items at any level cluster seemed to elicit and tap into a range of cognitive processes as discussed above, the levels appear more distinct when it comes to the type of information used by candidates to arrive at the correct answer. In line with the Aptis Listening Test specifications, the findings corroborate that A1 items largely require lexical recognition, A2 items tap almost exclusively into understanding factual information, B1 items require candidates to understand factual information as well as some interpretative meaning at the utterance level, and B2 items involve interpretative meaning at the utterance level and some understanding of meaning at discourse level.

Again, to probe this further, significance tests were conducted. As with the cognitive processes, we ran pairwise comparisons using Fisher exact tests and employing the Holm correction. The results of this are displayed in Table 18.

	1 Lexical recognition	2 Factual information	3 Interpretative meaning at the utterance level	4 Meaning at the discourse level
A1 – A2	0.00	0.00	0.50	1.00
A1 – B1	0.00	0.00	0.00	0.19
A1 – B2	0.00	0.21	0.00	0.00
A2 – B1	0.44	0.03	0.00	0.19
A2 – B2	0.60	0.00	0.00	0.00
B1 – B2	1.00	0.00	0.00	0.00

Table 18: p-Values of pairwise comparisons of information type ratios across CEFR levels (correct items only)

With few exceptions, this analysis shows a fairly clear picture of progression through the CEFR levels. We can see significant differences between the levels for most of the types of information. It appears that the Aptis Listening Test is more successful at targeting the desired type of information than the desired cognitive processes according to their test specifications. This, however, may either be due to the uniform test format used, or the fact that cognitive processes are more challenging to both elicit and code in stimulated recall protocols than type of information used. It may also be related to the difficulty of targeting or predicting the cognitive processes candidates are going to use in the test situation. In any case, the results show that the Aptis Listening Test is performing in accordance with the test specifications as far as the different types of information across CEFR levels are concerned.

6. DISCUSSION AND CONCLUSION

6.1 Cognitive processes and types of information used

The findings presented in this report indicate that the Aptis Listening Test successfully taps into the range of cognitive processes intended by the test developers. Test-takers answering the items correctly appear to employ a variety of processes as defined by Field (2013), which suggests a high level of cognitive validity in terms of the test's intended purpose as defined in the test specifications. The dataset further shows that the different item levels also elicit a range of processes each. However, the differences across the level clusters based on this dataset could be more distinct. No statistically significant progression from predominantly lower-level processing to increased proportions of higher-level processing could be found in pairwise comparisons of the items grouped by CEFR level. While it could be argued that the construct of listening, in terms of cognitive processing, is well-represented and adequately sampled from in the Aptis Listening Test overall, the differences between the CEFR levels could be more pronounced. The study found support for this in stimulated recall reports from candidates, which were carried out immediately after candidates had completed the items, and involved a replay of their eye-movements overlaid with the items' sound file as stimulus.

The fact that evidence of discourse construction could only be found in 60% of the correctly answered items at B2 level in the stimulated recall analysis, and that there was no statistically significant difference for the higher-order processes distinguishing this level from the lower levels, may suggest that there could be a better item type to tap into higher-order types of processing. The Aptis test designers may wish to consider alternatives to elicit discourse construction processing at B2, as it is specified as the key processing type to be elicited at this level.

The study also included an exploratory quantitative eye-movements analysis of candidates completing items on the Aptis Listening Test. The results of the linear mixed effects model analyses of the eye-tracking metrics show a difference in processing across CEFR levels and indicate that higher CEFR level items tend to elicit more processing on the responses, albeit it is not clear whether this relates to listening, reading, or test-taking processing, or, as is likely, a combination of the three.

While the listening test items were found to align with the test specifications in terms of intended cognitive processes only for some of the CEFR levels, in terms of the information targeted by the individual items this was different. In the stimulated recall dataset, a clear difference and progression regarding the types of information used by candidates to answer items correctly was observed. The test developers aim at testing lexical recognition at CEFR A1 level, factual information at A2 and B1 level, and interpretative meaning at the utterance level and meaning at discourse level at B2 level. Evidence for this was found in the data. The results appear to confirm that these information types are being successfully targeted by the items of different CEFR level clusters.

6.2 Unexpected findings: response order and reading ability

Apart from the results relating to cognitive processing and types of information used, which are mainly based on the stimulated recall protocol analysis, the eye-tracking analyses presented in this report also revealed two unexpected findings. One finding relates to the response order of items. The eye-tracking evidence suggests that the response order in multiple-choice listening tasks impacts the amount of focus on responses. It was found that responses higher up on the screen were looked at significantly longer than responses lower down, with a clear progression from the top to the bottom of the screen. While it is uncertain whether this difference also leads to differing amounts of cognitive processing of the individual responses, it is likely that responses presented higher up on the screen are more easily accessible to candidates and might therefore potentially impact item difficulty. However, this could be easily resolved by rearranging the responses on the screen.

The second unexpected finding of the eye-tracking analysis concerns the relative importance of reading and listening ability to answer items in listening tests. By analysing the eye-tracking metrics according to the reading and listening proficiency results of the candidates as measured by an additional test, tentative evidence was found to suggest that better readers read the responses more often than poorer readers. This is particularly interesting as, contrary to the reading component, the Aptis Listening Test only employs one item type (multiple-choice). However, it is not clear whether reading the responses more often aided more proficient readers in answering the items.

6.3 Potential and limitations of the methodology

Methodologically, the contributions and findings of this study have other implications beyond the present test being examined. Eye-tracking in conjunction with stimulated recall data has been illustrated as a valuable tool to investigate cognitive processes during listening test completion. Particularly the use of eye-movement recordings overlaid with sound files of listening test items as stimulus material for subsequent verbal reports seems to be a promising avenue of further research. In addition, the study has shown that eye-tracking itself can inform certain aspects of listening test-taking processes that other methods, such as stimulated recall, do not manage to capture. Finally, linear mixed effects modelling has been shown to be a useful analysis tool for controlling for factors that are challenging to handle in the experimental design or may not have even been considered as potentially confounding variables *a priori*.

The study has, however, also highlighted some of the limitations of the chosen methodology. Eye-tracking alone is of limited use to investigate which specific cognitive processes or type of information candidates use for answering multiple-choice items in listening tests. Also, the results show that stimulated recall does not allow for making automated lower-level processes such as input decoding visible. This means that the description of cognitive processes employed by candidates while taking the Aptis Listening Test may not be completely comprehensive.

Another limitation of the study concerns the nature of the candidate sample. Despite best efforts to recruit a balanced participant sample for the study, ranging from very low to relatively high proficiency second language learners of English, the sample could be criticised as skewed towards more proficient learners. Thus, the study was unfortunately not able to reveal whether less proficient learners use different cognitive processes or types of information at the different question levels. Further research would need to consider a wider range of proficiency levels and a potentially larger sample size of candidates. However, given the time-consuming nature of the approach adopted, the sample size of the present study appears substantial.

6.4 Areas for future research

Despite these limitations, the study revealed important areas for future research. Using eye-tracking to investigate the relationship between listening and reading ability and the propensity to use information from either channel in answering listening test items could prove fruitful, as the current study's findings appear to underline the important role of reading in answering listening test items. In this respect, it would be insightful to replicate the study's design using other task types than multiple-choice.

The findings regarding the impact of response order on visit duration of individual responses suggests that further work should also be carried out in experimenting with different screen layout arrangements for multiple-choice tasks, so that each response gets roughly equal attention. It is hypothesised from the results of the present study that test-taking strategies, in particular the ability or awareness to carefully read all responses, might interact with the intended construct and item difficulty, causing potential construct-irrelevant variance.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Baayen, R. H., Davidson, D. J. & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Badger, R. & Yan, X. (2012). The use of tactics and strategies by Chinese students in the listening component of IELTS. In L. Taylor & C. J. Weir (Eds.), *IELTS Collected Papers 2: Research in reading and listening assessment* (pp. 454–486). Cambridge: Cambridge University Press.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, *30*(4), 441–465. http://doi.org/10.1177/0265532212473244

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

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

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language. *Cognitive Psychology*, *6*(1), 84–107.

Crawley, M. J. (2007). Mixed-effects models. *The R Book,* 681–714. Wiley Publishing.

Ericsson, K. A. & Simon, H. A. (1987). Verbal reports on thinking. In C. Faerch & G. Kasper (Eds.), *Introspection in second language research* (pp. 24–53). Philadelphia: Multilingual Matters.

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, Massachusetts: MIT Press.

Field, J. (2008). *Listening in the language classroom*. Cambridge: Cambridge University Press.

Field, J. (2012). The cognitive validity of the lecture-based question in the IELTS listening paper. In L. Taylor & C. J. Weir (Eds.), *IELTS Collected Papers 2: Research in reading and listening assessment* (pp. 391–453). Cambridge: Cambridge University Press.

Field, J. (2013). Cognitive validity. In L. Taylor & A. Geranpayeh (Eds.), *Examining listening* (pp. 77–151). Cambridge: Cambridge University Press.

Field, J. (2015). The effects of single and double play upon test outcomes and cognitive processing. *ARAGs Research Reports Online*. London. Retrieved from www.britishcouncil.org/exam/aptis/research/publications

Gass, S. & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Gelman, A. & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Godfroid, A. & Spino, L. A. (2015). Reconceptualizing Reactivity of Think-Alouds and Eye Tracking: Absence of Evidence Is Not Evidence of Absence. *Language Learning*, *65*(4), 896–928.

Green, A. (1998). *Verbal protocol analysis in language testing research*. Cambridge: Cambridge University Press.

Harding, L. (2011). *Accent and listening assessment. Language Testing and Evaluation Volume 21*. Frankfurt am Main: Peter Lang.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

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

Kuznetsova, A., Brockhoff, P. B. & Bojesen Christensen, R. H. (2016). lmerTest: Tests in linear mixed effects models. R package version 2.0-3.0. Retrieved 16 August 2016, from http://cran.r-project.org/package=lmerTest

McCray, G. (2013). *Statistical modelling of cognitive processing in reading comprehension in the context of language testing*. Unpublished PhD thesis: Lancaster University, UK.

McCray, G., Alderson, J. C. & Brunfaut, T. (2012). *Validity in reading comprehension items: Triangulation of eye-tracking and stimulated recall data*. Paper presented at the EALTA conference: University of Innsbruck, Austria.

McCray, G. & Brunfaut, T. (2016). Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. *Language Testing*. http://doi.org/10.1177/0265532216677105

Nakazawa, M. (2015). fmsb: Functions for Medical Statistics Book with some Demographic Data. R package version 0.5.2.

O'Sullivan, B. (2011). Language testing. In J. Simpson (Ed.), *Routledge Handbook of Applied Linguistics* (pp. 259–273). Oxford: Routledge.

O'Sullivan, B. & Dunlea, J. (2015). *Aptis general technical manual*. Retrieved from https://www.britishcouncil.org/sites/default/files/aptis_general_technical_manual_v-1.0.pdf

O'Sullivan, B. & Weir, C. J. (2011). Language testing and validation. In B. O'Sullivan (Ed.), *Language testing: Theory and practice* (pp. 13–32). Oxford: Palgrave.

R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved 16 August 2016, from http://www.r-project.org

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506. http://doi.org/10.1080/17470210902816461

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M. & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Taylor, L. & Geranpayeh, A. (Eds.). (2013). *Examining listening*. Cambridge: Cambridge University Press.

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, *40*(3), 191. http://doi.org/10.1017/S0261444807004338

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

Winke, P. & Lim, H. (2014). The effects of testwiseness and test-taking anxiety on L2 listening test performance: A visual (eye-tracking) and attentional investigation. *IELTS Research Reports Online Series*, (3), 1–30.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *arXiv Preprint arXiv:1308.5499*.

British Council Assessment Research Awards and Grants

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

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

LOOKING INTO LISTENING:
Using eye-tracking to establish the cognitive validity of the Aptis Listening Test

AR-G/2017/3

Franz Holzknecht, Kathrin Eberharter
Benjamin Kremmel, Gareth McCray,
Matthias Zehentner, Eva Konrad,
Carol Spöttl

ARAGs RESEARCH REPORTS ONLINE

ISSN 2057-5203