GEPT®
LTTC 全民英檢

Language unfolds worlds.
Testing sets standards.

LTTC-GEPT Research Reports RG-07

Researching the Cognitive Validity of GEPT High-Intermediate and Advanced Reading An Eye Tracking and Stimulated Recall Study

Stephen Bax
Sathena H. C. Chan

LTTC® 財團法人語言訓練測驗中心
THE LANGUAGE TRAINING & TESTING CENTER

Researching the Cognitive Validity of GEPT High-Intermediate and Advanced Reading An Eye Tracking and Stimulated Recall Study

LTTC-GEPT Research Reports
RG-07

Stephen Bax
Sathena H. C. Chan

This study was funded and supported by the Language Training & Testing Center (LTTC) under the LTTC-GEPT Research Grants Program 2014-2015

LTTC-GEPT Research Reports RG-07
Researching the Cognitive Validity of GEPT High-Intermediate and Advanced Reading: An Eye Tracking and Stimulated Recall Study

Published by The Language Training and Testing Center
No.170, Sec. 2, Xinhai Rd., Daan Dist., Taipei, 10663 Taiwan (R.O.C)

First published July 2016

Foreword

We have great pleasure in publishing this report: *LTTC-GEPT Research Reports RG-07*. The study described in this report was funded by the 2014-2015 LTTC-GEPT Research Grants. Headed by Professor Stephen Bax of Open University, UK, the study investigated the cognitive validity of the GEPT High-Intermediate and Advanced Reading Tests using eye tracking technology and stimulated recall procedures. The study provides empirical evidence for the cognitive validity of the GEPT High-Intermediate and Advanced Reading Tests.

The GEPT, developed more than a decade ago by the LTTC to serve as a fair and reliable testing system for EFL learners, has gained wide recognition in Taiwan and abroad. It has generated positive washback effects on English education in Taiwan. As the GEPT has successfully reached out to the international academic community with remarkable success over the years, numerous studies and research projects on GEPT-related subjects have been conducted and published as technical monographs, conference papers, and refereed articles in books and journals. In view of the growing scholarly attention on the GEPT, and in order to assist external researchers to conduct quality research on topics related to the test, the LTTC has set up the LTTC-GEPT Research Grants Program, which offers funding to outstanding research projects.

The annual call for research proposals is publicized every October, attracting proposals from all over the world. A review board, which comprises scholars and experts in English language teaching and testing from Taiwan and abroad, evaluates the research proposals in terms of the following criteria:

- the relevance to identified areas of research
- the benefit of the research outcomes to the GEPT
- the theoretical framework, aims and objectives, and methodology of the proposed research
- the qualifications and experience of the research team
- the capability of the research outcomes to be presented at international conferences and published in journals
- the timeline and cost effectiveness of the proposed research

Complete and up-to-date information about the GEPT is available at https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT.htm. Full research reports can be downloaded at https://www.lttc.ntu.edu.tw/lttc-gept-grants.htm.

We believe that with the further contributions from the external research community, the GEPT will continue to refine its quality and achieve wider recognition at home and overseas.

Hsien-hao Liao
Executive Director
LTTC

Author Biodata

Professor Stephen Bax is Professor of Modern Languages and Linguistics at the Open University in the UK. In 2014 he was awarded the *TESOL Distinguished Researcher Award* for his research on eye tracking, presented at the TESOL convention in Portland, Oregon. He has a PhD from the University of Kent in the area of Discourse, and MLitt and MSc in Applied Linguistics from the University of Edinburgh. His research interests include the use of eye tracking to research reading, the use of computers in language learning (CALL), the use of computers in language testing (CALT), and areas of discourse including Computer Mediated Discourse Analysis. Most recently he has been researching the role of vocabulary in texts using his online tool *textinspector.com*. For the BBC World Service website he has produced numerous internet-based learning resources, including a major interactive language learning series "Ten Days" for Latin America and internationally, as well as numerous other interactive language learning modules.

Dr. Sathena Chan is Lecturer in Language Assessment at CRELLA, the Centre for Research in English Language Learning and Assessment at the University of Bedfordshire. She holds a PhD from University of Bedfordshire, and an MA in Language Testing and Programme Evaluation from the University of Essex. Her major areas of research include integrated reading-into-writing assessment, test development and validation, cognitive processing of language use, criterial analysis of written performance, and rating scale development. She has extensive experience in statistics analyses for language assessment and second language learning. Since joining CRELLA as a full-time member in 2013, she has been actively involved in different test development and validation projects for examination boards and educational organisations in the UK and worldwide. She has presented her research at a number of national and international conferences, such as LTRC, the ALTE International Conference, the EALTA Conference and the Language Testing Forum (LTF).

摘要

◆ **研究團隊與研究目的**

本研究由英國公開大學（Open University）Stephen Bax 教授與貝德福德大學（University of Bedfordshire）Sathena Chan 博士共同主持，目的是探討考生於全民英檢中高級、高級閱讀測驗的認知過程，研究結果為全民英檢中高級、高級閱讀測驗提供進一步的效度證據。

◆ **研究問題**

1. 探索全民英檢中高級閱讀測驗各部份引導的認知歷程。
2. 探索全民英檢高級閱讀測驗各部份所採取的認知歷程。
3. 兩級數閱讀測驗引導考生應用認知歷程的向度是否符合所預期？

◆ **研究方法摘要**

1. 本研究採取眼動追蹤科技（eye-tracking technology）紀錄考生於參與全民英檢中高級、高級閱讀測驗的認知歷程，並搭配刺激回憶法（stimulated recall）進行資料收集。
2. 受試者為 24 位正於英國修讀不同學科碩士學位的台灣學生，IELTS 平均成績為 6.58（標準差為.54， 最高分 8 級分，最低分 6 級分，差距為 2）。
3. 資料分析主要分為兩部份：(1)受試者測驗時的眼動軌跡，與(2)試後問卷調查與訪談。受試者考試中眼動軌跡使用 *Tobii X2 Eye Tracker* 記錄。測驗結束後，受試者立即填寫一份認知歷程問卷。部份受試者受邀進行試後的訪談，訪談時受試者觀看自己眼動軌跡的紀錄影片，並向研究團隊說明當時的閱讀行為。

◆ **研究結果摘要**

1. 全民英檢中高級閱讀測驗引導考生應用的認知歷程，包含：
   - Word recognition
   - Lexical access
   - Syntactic parsing
   - Inferencing
   - Establishing propositional meaning at clause and sentence level
   - Integrating information across sentences
   - Creating a text level structure
2. 全民英檢高級閱讀測驗引導的認知歷程包含前述中高級的七項，以及最高階的 Integrating information across texts。（Khalifa and Weir (2009)定義八個層次認知歷程中的前七個層次）
3. 兩級數閱讀測驗所引導的認知歷程與所預期相符。此外，答對與答錯考生的眼動軌跡不同，而答對的受試者於答題的過程中採取較多高層次的認知策略。

ABSTRACT

It is important for any language test to establish its *cognitive validity* in order to ensure that the test elicits from test takers those cognitive processes which correspond to the processes which they would normally employ in the target real-life context (Weir 2005). This study investigates the cognitive validity of the GEPT Reading Test at two levels, High-intermediate (CEFR B2) and Advanced (CEFR C1), using innovative eye-tracking technology and detailed stimulated recall interviews and surveys.

Representative reading items were carefully selected from across all parts of the GEPT High-Intermediate Level Reading Test and the GEPT Advanced Level Reading Test. Taiwanese students (n=24) studying Masters level programmes at British universities were asked to complete the test items on a computer, while the *Tobii X2 Eye Tracker* was used to track their gaze behaviour during completion of the test items. Immediately after they had completed each individual part, they were asked to report the cognitive process they employed by using a Reading Process Checklist, and a further (n=8) then participated in a detailed stimulated recall interview while viewing video footage of their gaze patterns.

Taking into account all these sources of data, it was found that the High-Intermediate section of the GEPT test successfully elicited and tested an appropriate range of lower and higher cognitive processes, as defined in Khalifa and Weir (2009). It was also concluded that the Advanced sections of the test elicited the same set of cognitive processes as the High-Intermediate test, with the addition in the final section of the most difficult of all in Khalifa and Weir's scheme.

In summary, it is apparent that the two elements of the GEPT test which were researched in this project were successful in requiring of candidates the range of cognitive processing activity commensurate with High-Intermediate and Advanced reading levels respectively, which is an important element in establishing the cognitive validity of the GEPT test.

Table of Contents

Table of Tables

Table of Figures

1. AIMS AND OBJECTIVES

Research demonstrating the cognitive validity of the GEPT Reading Test is critical, as such evidence is essential if the test is to gain greater acceptance from educational institutions as a measure of international students' academic English proficiency. Since the 1990s it has been argued that language tests assessing complex cognitive constructs should establish their *cognitive validity* (Glaser 1991, Baxter & Glaser, 1998) since cognitive interpretative claims are "not foregone conclusions, [but] need to be warranted conceptually and empirically" (Ruiz-Primo et al. 2001:100). The primary aim of this study is therefore to assist in the project to validate the cognitive validity of the GEPT Reading Test at two levels: High-intermediate (B2) and Advanced (C1).

2. LITERATURE REVIEW AND RESEARCH QUESTIONS

2.1 Theoretical support for cognitive validation

Cognitive validity is a fundamental component in Weir's (2005) socio-cognitive validation framework for language tests, which marks the first systematic attempt at providing language testing stakeholders, such as test developers, test takers and test score users (e.g. universities and teachers) with a coherent and accessible methodology for test development and validation. This framework conceptualises the test validation process by identifying different types of validity evidence which need to be collected at different stages, i.e. the a priori and a posteriori stages, of test development and validation (Geranpayeh & Taylor (eds), 2013: 27). The framework covers five components of test validity: (1) context validity, (2) cognitive validity, (3) scoring validity, (4) consequential validity and (5) criterion-related validity. The proposed study focuses on the cognitive validity of the GEPT Reading Test. Cognitive validity (Glaser, 1991) addresses the extent to which a test elicits from test takers cognitive processes that correspond to the processes which they would normally employ in a real-life context. Khalifa and Weir (2009) decomposed the cognitive processes in reading (see *Table 1* below) and investigated how these processes are operationalised by the Cambridge reading examinations at different levels, considering that such a process was necessary to demonstrate the examinations' validity.

Table 1. Reading processes identified by Khalifa and Weir's (2009) model

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

In a similar vein, Wu (2014) investigated the cognitive processes involved in the GEPT Reading Test at the Intermediate (B1) and High-Intermediate (B2) levels through expert judgment and a test taker cognitive process checklist. The results of the expert judgment provided a useful indication of the set of target cognitive processes which may arguably be elicited from the test takers. However, empirical evidence is required to reveal what cognitive processes are actually employed by the test-takers under test conditions. Similarly, the checklist only assisted the test takers to report the perceived processes rather than the actual processes employed. The importance of collecting such evidence of the actual processes employed, especially for receptive language skills such as reading, has been established in the literature. For example, Alderson (2007) states that:

> *There was no theory of comprehension that could be used to identify the mental operations that a reader or listener has to engage in at the different levels of the CEFR. Yet such a theory is essential if one is to begin to identify the development of so-called receptive abilities in CEFR terms* (p. 661).

2.2 Use of eye-tracking technology to investigate cognitive processes

To elucidate cognitive processes in reading tests, most studies have used think-aloud protocols and interview (e.g. Cohen & Upton, 2007), questionnaires/checklists (e.g. Wu, 2014), and expert judgment (e.g. Wu, 2014). However, these methods are limited, especially when used in isolation. For example, expert judgment only reveals the set of target cognitive processes from the perspective of a group of experts in the field whereas think-aloud procedures arguably disrupt the processes under investigation (Cooper & Holzman, 1983; Russo, Johnson & Stephens, 1989).

Recently an innovative method to supplement these existing methods has become available through developments in non-intrusive eye tracking technology (Eger, Ball, Stevens & Dodd, 2007). In recent research projects funded respectively by Cambridge ESOL, the British Council and IELTS, Bax and colleagues at CRELLA, the University of Bedfordshire, have made innovative use of eye tracking equipment to investigate test-takers' onscreen reading (Bax, 2012, 2013a, 2013b). These screenshots show examples of gaze patterns in reading during these projects:

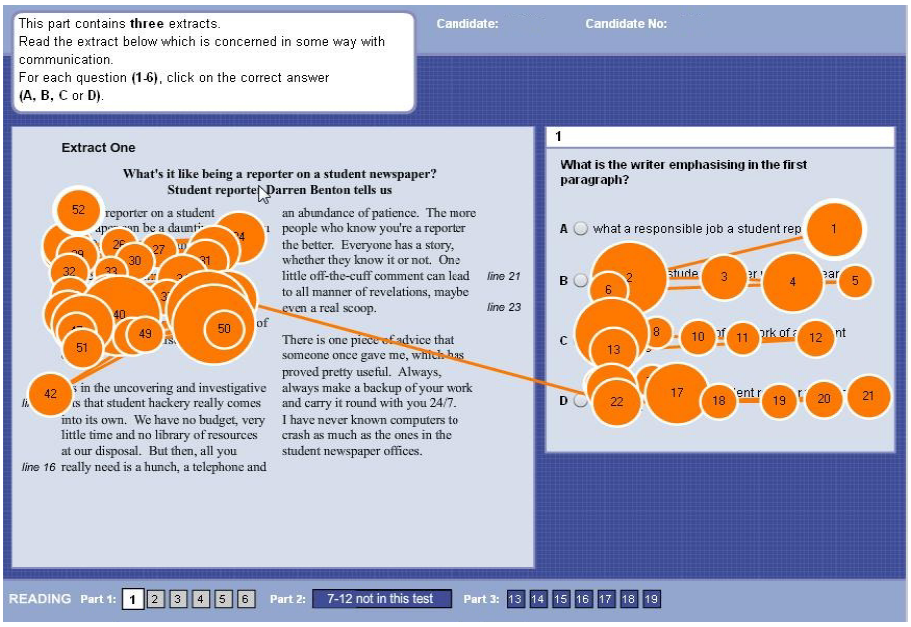
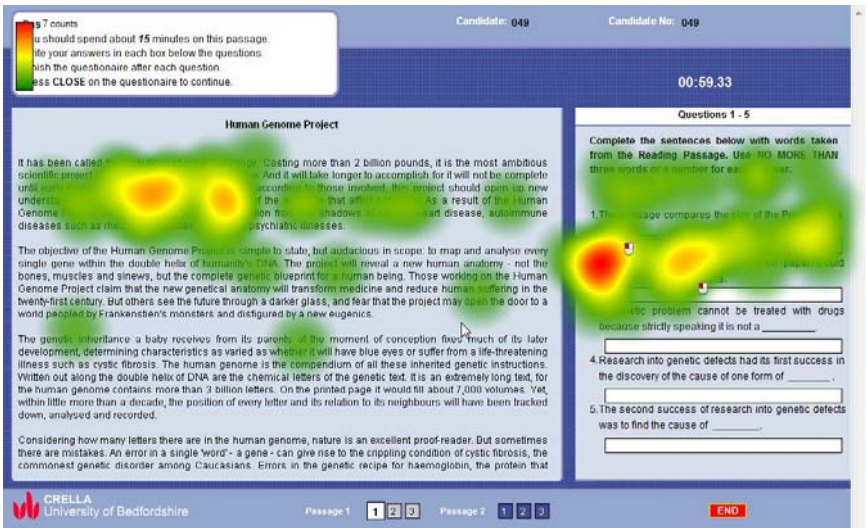


Figure 1. Gazeplot: eye movements during a reading test



Heatmap: user's fixation patterns

These studies show that eye tracking, when used in conjunction with other tools such as retrospective reports, can offer effective, non-intrusive insights into cognitive processing under test conditions, when used in conjunction with other methods. In particular the research reported in Bax (2013b) in *Language Testing* – the first study in the journal which uses eye tracking in this way – offers compelling evidence of the value of eye tracking for researching cognitive validity in reading tests. Since eye tracking data can also identify disparities between participants' self reports and their actual behaviour as revealed by eye movements, the technology – when combined with verbal reports – promises more reliable data than such reports alone. In addition, it generates recordings of users' eye movements which are valuable in prompting participants' recall.

2.3 Research Questions

In the light of the studies discussed above the following three research questions were identified in relation to the GEPT examination:

1. What cognitive processes are elicited by different sections of the GEPT High-Intermediate Level Reading Test?

2. What cognitive processes are employed by test-takers on different sections of the GEPT Advanced Level Reading Test?

3. To what extent and in what ways do the cognitive processes elicited at the two levels match the cognitive processes anticipated in reading tests at these levels?

These questions will be revisited after all the data, from participants' performance, eye tracking, self report (reading processing checklist) and interviews, have been considered (see *5.1 Research Questions revisited* on page 28 below).

3. METHODOLOGY

3.1 Tasks

Reading items were carefully selected from across all parts of the GEPT High-Intermediate Level Reading Test and the GEPT Advanced Level Reading Test except for one[1] (for details see *Table 2*). A testlet with all the selected items was created using Adobe Acrobat DC to facilitate the eye tracking study, with careful attention paid to making the test experience as close as possible to normal GEPT test contexts. The time allocation for each item was calculated based as a proportion of the original time allocation in the original test.

[1] Part 1 (Sentence Completion) of the GEPT High-Intermediate Reading Test was excluded because this section targets lexico-grammatical knowledge at sentence level rather than reading comprehension.

Table 2. Selected reading items from the GEPT Reading Test for the investigation

Level	Items in GEPT test	Test item numbers in the current research project	Items identified for detailed analysis	Cognitive processes involved (minimum) *(see Table 1, page 1)*
High-Intermediate Level (14 mins)	Part 2 Cloze (n=7) • Q16-22 (MCQ)	1-7	3	-Syntactic parsing
	Part 3 Reading Comprehension (n=7) • Q37-38: Graph (MCQ) • Q46-50: Article (MCQ)	8-9 10-14	9 13 14	-Inferencing -Establishing propositional meaning at clause and sentence level -Integrating information across sentences -Creating a text level structure
Advanced Level (25 mins)	Part 1 Careful reading (15 mins) (n=6) • Q15-20 (Summary - Fill in the blanks)	15-20	19	-Word recognition -Lexical access -Integrating information across sentences -Inferencing -Creating a text level structure
	Part 2 Skimming & Scanning (10 mins) (n=10) • Q21-26 (Headings matching) • Q33-36 (Which text)	21-26 27-30	21 30	-Word recognition -Lexical access -Integrating information across sentences -Inferencing -Integrating information across texts
Total: 39 minutes	**30 test items**	**30 test items**	**7 items**	

The selection of items in this way allows for full coverage of question and response types, a variety of text lengths, and a range of potential cognitive processes so as to represent as accurately as possible the experience of both tests. Participants' performance was then marked by the researchers using the marking scheme provided by the LTTC. For the second part of the analysis, of the eye tracking gaze data, 7 representative items were selected for detailed investigation, as listed in *Table 2*, in ways to be discussed below.

3.2 Participants

Data was collected from 24 Taiwanese students studying Masters level programmes at British

universities. Their mean IELTS reading score was 6.58 with a standard deviation of 0.54 (max = 8, min = 6, range = 2). This means that in terms of level they were fairly homogenous as a group, with relatively little to divide them. They were recruited from various disciplines including Business, Education, Translation and Law to ensure a wide spectrum of reading approaches.

3.3 Eye-tracking technology

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

3.4 Reading Processing Checklist

A Reading Processing Checklist was developed based on Khalifa and Weir's (2009) model on reading processes and Wu's (2014) Cognitive Processing Checklist to assist participants to report the cognitive processes they employed immediately after they have completed each individual part of the GEPT Reading Tests (see *Appendix 1*).

3.5 Data collection

Data was collected on a one-to-one basis. Participants were asked to complete the GEPT test (see *Table 2*) on a computer. The *Tobii X2 Eye Tracker* was used to track the test takers' gaze behaviour on all the test items. Immediately after they had completed each individual part, they were asked to report the cognitive process they employed by using the Reading Process Checklist. A selected sample of the participants (33%, n=8) then participated in a stimulated recall interview. The key stages of the data collection included:

(1) Participants completed all personal information forms and consent forms;
(2) Researcher calibrated individual eye fixations and saccades for each participant using the device's calibration tool, which could identify each person's individual pattern of gaze and saccade behaviour and ensured the accuracy of the subsequent tracking of their reading during the test;
(3) Participants watched a short video tutorial, explaining each aspect of the process they were about to follow;
(4) Individual participants completed the GEPT High-Intermediate Level Reading Test and the GEPT Advanced Level Reading Test on screen;
(5) Immediately after completing each individual part, participants reported the cognitive process employed by using the Reading Process Checklist.
(6) A sample of participants then completed a stimulated recall interview procedure; participants each viewed the video footage of their own test, observing their eye movements represented on the screen. Participants described their reading behaviour during the video. The video was slowed, stopped and/or rewound at their request to allow them to view and comment freely.

3.6 Data analysis

3.6.1 Eye tracking data

The eye tracking data was analysed both qualitatively and quantitatively. Qualitatively, the segments (i.e. Areas of Interest) of the question and reading text(s) which test takers read to complete each individual test item were analysed. The cognitive demands (e.g. at the levels of lexical, single sentence, multiple sentences, single text or multiple texts) of each selected reading item were then coded using Khalifa and Weir's (2009) model, as described below. Findings are discussed in terms of differences between successful candidates on each item and unsuccessful candidates.

Quantitatively, numerical data on each test taker's eye movements, in terms of *Fixation Duration* (the length of time a reader fixated on a section of the text), *Fixation Count* (the number of times a reader fixated on a section of the text), *Visit Duration* (the length of time a reader remained on a section of the text), and *Visit Count* (the number of visits a reader made to a section of the text), were generated by the eye-tracking software. The particular sections of text selected for analysis varied from item to item (and are detailed in the Results section below), and included text as small as a phrase, a correct or incorrect answer in a Multiple choice question, or text as large as a whole passage. The eye-movement data were compared between the successful and unsuccessful test takers (i.e. those who answered the item correctly and those who answered the item incorrectly) using descriptive statistics.

The non-parametric Mann-Whitney U tests were also conducted to examine the statistical significance of any differences of the eye movements between the successful and unsuccessful students on an item-by-item basis, when the sample size of both groups exceeded n=5. It is recognised that the Mann-Whitney U test can be used with such small samples (see e.g. Sheskin 2003, Hinton 1995, and in particular the example in Wood, Fletcher & Hughes, 1986, page 188). However, the sample size of one of the two groups was often very small, and therefore the results of the inferential statistics should be interpreted with caution. In addition, close analysis was applied in more qualitative mode to the detailed gaze patterns of each participant, via the heat map and other tools available in the software, in order to tease out other important patterns which might be missed by purely quantitative approaches.

3.6.2 Self-report and Stimulated recall data

As an important corollary to the eye tracking research, we then obtained reports from participants on their reading activities, via the Reading Process Checklist and Verbal protocols, in ways set out in section 3.5 above. In order to obtain a more nuanced and fine-grained understanding of how test-takers' reading and cognitive processing was related to test scores (so as to answer all three of our research questions), participants were sorted into four groups according to levels of performance (i.e. low, low-medium, medium-high, high scoring groups - see *Table 4* in the Results section). Frequencies of participants' responses in the Reading Process Checklist in relation to reading goals, levels of processing and levels of

comprehension were then calculated to provide an overall analysis of the processes elicited by different parts of the two GEPT Reading Tests. Verbal protocols elicited in the stimulated recall interview were transcribed to supplement the self-report processing data. After that, the data was compared in two strands: (a) responses by different scoring groups (see *Table 4* in the Results section), and (b) responses from the items of the GEPT High-Intermediate Level Reading Test (B2) and the GEPT Advanced Level Reading Test (C1).

4. RESULTS

4.1 Performance

Table 3. Participants' performance

Level	Part	Format	Mean	SD	Max	Min
			%	**%**	**%**	**%**
High-Intermediate Level (B2)	Part 1 Cloze	MCQ	74.40	13.03	100.00	57.14
	Part 2 Reading Comprehension	MCQ	71.43	23.33	100.00	28.57
Advanced Level (C1)	Part 3 Summary	Fill in the blanks	28.13	20.81	75.00	0
	Part 4 Skimming & Scanning	Headings matching	40.42	19.03	80.00	10.00
		Total	53.13	14.82	78.33	30.00

It will be noted from *Table 3* above that participants performed relatively well on the High-Intermediate section of the test overall, and relatively poorly on the Advanced section, in particular on the Summary. At interview they noted that the Summary was particularly difficult.

For purposes of the eye tracking analysis candidates were divided according to their correct or incorrect answers on a question-by-question basis. There were three reasons for not dividing them into groups according to their IELTS score: in the first place the differences in their IELTS scores were too small to differentiate them with confidence; secondly the IELTS reading scores specifically were not available; thirdly, and most importantly, the key issue in our research was whether a candidate had answered a particular test item correctly and why, so it was essential to differentiate them on a question-by-question basis.

For the purpose of interpreting the self-report and interview data, a finer distinction was possible, with participants categorised into either the low-scoring (achieving 0-25% of the total score), low-medium (26-50%), medium-high (51-75%) or high-scoring (76-100%) group, depending on their performance on each Part (see *Table 4*).

Table 4. Grouping based on participants' performance

	Part 1 Cloze	Part 2 Reading Comprehension	Part 3 Summary	Part 4 Skimming & Scanning
Low (0-25%)	0	1	15	6
Low-medium (26-50%)	0	5	7	14
Medium-High (51-75%)	15	8	2	2
High (76-100%)	9	10	0	2

4.2 Eye-tracking data

In this section we now present the results from the eye tracking section of the research. In this section we give results for each of the items in their own terms. After reporting on the self-report and interview data we then summarise the implications for each of our Research Questions in *5.1 Research Questions revisited,* page 28 below.

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

High-Intermediate Items

Item 3

(High-Intermediate Level, Part 2 Cloze)

Item 3 was selected because it offers an interesting focus on grammar, and was answered correctly by a relatively small number of candidates (8 correct and 16 incorrect). The item expected candidates to connect two clauses with the correct subordinating conjunction *'in that'*, as follows:

> *Item 3: Children's triathlons also differ ..... (3)..... they place less emphasis on competition than on participation.*
>
> *A. much from*
>
> *B. in that (correct answer)*
>
> *C. with which*
>
> *D. other than*

This related therefore to the lower level process termed by Khalifa and Weir *"syntactic parsing"* (see *Table 1*). To answer this item correctly candidates were required a) to read the whole sentence carefully, and then b) to use their grammatical knowledge to distinguish between the correct and incorrect options. For this reason we would be unlikely to see any clear differences in the eye tracking gaze data between candidates' activity, because most of their activity while answering this item would not involve gaze movements, but would consist largely of purely mental activity inaccessible to eye tracking equipment.

Nonetheless, we attempted to compare the successful students' gaze data with gaze data from unsuccessful students on three Areas of Interest, namely:

-the target sentence as a whole
-the correct response only (to see if successful or unsuccessful students focussed on it more intensively)
-the incorrect responses (to see if successful or unsuccessful students focussed on them more intensively)

Gaze data was analysed for *Fixation Duration*, *Fixation Count*, *Visit Duration* and *Visit Count*. To ensure complete accuracy of analysis those candidates were excluded whose gaze date on this item was in any way unusual, defective or unclear, so data from a total of 4 successful and 13 unsuccessful candidates' gaze activity was closely analysed both quantitatively and qualitatively.

Results for Item 3

The data was analysed qualitatively and quantitatively, although the use of inferential statistics was not possible because of the small number of correct candidates analysed. As expected, for reasons noted above, candidates did not show notable differences in eye gaze activity for this item, since presumably their activity in answering this item was almost entirely cognitive, with relatively little ocular activity. Furthermore, owing to the nature of the item, successful candidates were presumably distinguished overwhelmingly in terms of correct (invisible) grammatical parsing, not by any differences in observable eye activity.

For these reasons it was not possible with this item to identify clear differences in gaze activity between successful and unsuccessful candidates. However, there did appear to be a modest difference in terms of overall attention paid to the text, with the successful candidates appearing to fixate more extensively and frequently on the target text, and also on the possible responses (correct and incorrect in terms of all the measures - Fixation Duration, Fixation Count, Visit Duration and Visit Count). As can be seen in Table 5, successful candidates outranked unsuccessful candidates on every measure, indicating that they attended more closely to the target text, the only exceptions (highlighted) being on the Incorrect options in the MCQ item, in terms of Mean Fixation Duration, Mean Fixation Count and Mean Visit Count. This could be explained by the fact that unsuccessful candidates are more likely to fixate more intensively and visit more frequently the Incorrect options which they are about to choose.

Table 5. Gaze data results for Item 3

Item 3: Target sentence in text								
	Mean Fixation Duration (secs)		Mean Fixation count (number)		Mean Visit duration (secs)		Mean Visit count (number)	
Successful candidates (S) (n=4) or unsuccessful (n=13) (U)	S	U	S	U	S	U	S	U
Mean	13.80	8.34	76.50	48.00	1.10	0.73	19.50	17.62
SD	7.39	4.61	37.14	25.56	0.56	0.35	9.00	10.10
Item 3: Correct option in the MCQ								
	Mean Fixation Duration (secs)		Mean Fixation count (number)		Mean Visit duration (secs)		Mean Visit count (number)	
Successful candidates (S) (n=4) or unsuccessful (n=13) (U)	S	U	S	U	S	U	S	U
Mean	2.62	2.47	12.25	12.00	0.47	0.44	7.25	6.67
SD	0.82	1.97	4.92	8.31	0.27	0.25	2.22	3.63
Item 3: Incorrect options in the MCQ								
	Mean Fixation Duration (secs)		Mean Fixation count (number)		Mean Visit duration (secs)		Mean Visit count (number)	
Successful candidates (S) (n=4) or unsuccessful (n=13) (U)	S	U	S	U	S	U	S	U
Mean	***7.75***	***10.91***	***32.05***	***33.08***	1.08	0.86	***15.75***	***17.55***
SD	4.11	6.00	**15.44**	**16.12**	0.51	0.34	**9.54**	**8.57**

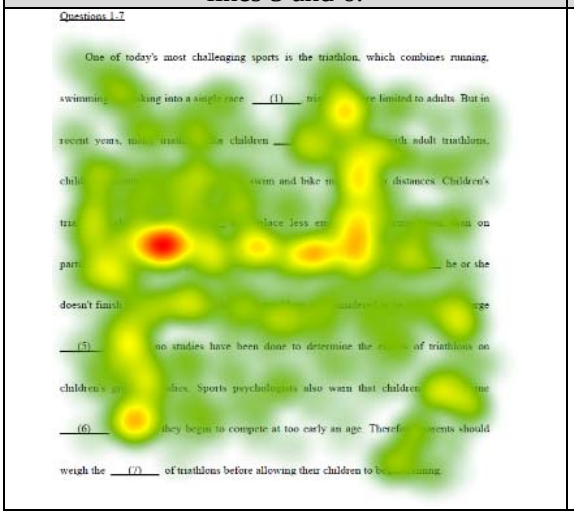
Heatmap of gaze behaviour of a successful candidate. Note the intense concentration on lines 5 and 6.	Heatmap of unsuccessful candidate. Note relatively low concentration on lines 5 and 6.
	

Figure 2

Figure 2 offers an interesting illustration of the trend discussed above with Item 3, in that the

successful candidate, whose gaze patterns are presented as the heatmap on the left, clearly focussed far more intensively on all parts of the target sentence around the item (five lines down, starting with the word 'triathlons') than did the unsuccessful candidate on the right, whose focus on that area was patchy.

In summary, then, the nature of Item 3 with its focus on grammar meant that it was not possible with any certainty to differentiate definitively between successful and unsuccessful readers in terms of gaze information. Nonetheless, there was some evidence that successful candidates fixated on and visited the target text and the test items more than unsuccessful candidates, and were rewarded for this. It is probable (given the item type) that this tells us more about effort and concentration than about cognitive processing, but it is interesting nonetheless.

Item 9
(High-Intermediate Level, Part 3 Reading Comprehension)

This item was chosen because it related to the reading of the graph, and so represented an interesting variation on the kinds of reading tested by other items. Readers had to look at the graph and answer this question:

> *Which countries had approximately the same number of Internet users in the fourth quarter as in the first?*
> *A. Australia and Spain*
> *B. Spain and Sweden (correct answer)*
> *C. Mexico and Sweden*
> *D. Australia and Mexico*

Since readers were attempting also to answer a second question relating to the graph (item 8) it was not possible within the eye tracking gaze data to separate out the gaze pattern activity relating specifically to this item as they looked at the graph, so gaze data was gathered only from the item itself, relating to Fixation and Visits on the test item, the correct option and the incorrect options. It was interesting that almost all candidates identified the correct answers (see *Table 3* above) - in fact 19 of the 24 candidates (79%) were correct. This also meant that data on unsuccessful candidates was very limited.

Possibly for this reason, there were no notable differences seen between gaze patterns of successful and unsuccessful candidates on any of the areas of interest for this item. However, one finding from the qualitative analysis was interesting. The candidates who were incorrect on this item spent relatively little time on the item itself. As can be seen from **Figure 3**, the image on the left, whereas they focused heavily on item 8 at the top of the image, their focus on item 9 was relatively scant. By contrast we can see the way which one successful candidate (**Figure 3**, on the right) focussed as intently on item 9 as he had on item 8, and in fact got both correct. This might suggest that because the item appeared easy, with little inherent cognitive

challenge, unsuccessful candidates underestimated it and did not distinguish closely enough between the distractors.

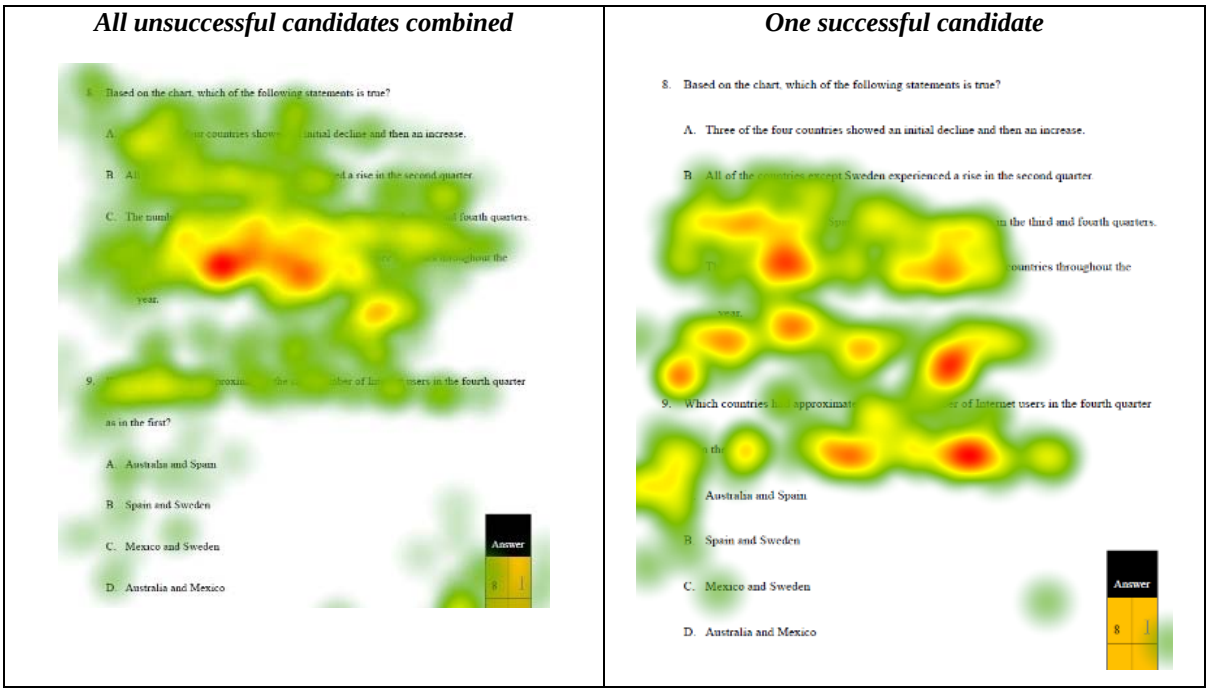


Figure 3. Gaze data heatmaps for Item 9

Item 13

(High-Intermediate Level, Reading comprehension, Article with MCQs)

From this section of the High-Intermediate Level test we identified two test items to analyse in detail, one which required close attention to a specific section of the text (Item 13) and one which required understanding of the whole text (Item 14). Investigating these items was therefore important in our attempt to understand the whole range of cognitive processes which Khalifa and Weir identified (see *Table 1*, page 1), since they required respectively *"Establishing propositional meaning at clause and sentence level"* and *"Creating a text level structure"*. Item 13 was as follows:

> *What was van Meegeren originally accused of doing?*
> *A. Assisting foreigners to obtain a national treasure (correct answer)*
> *B. Trading fake paintings for special privileges*
> *C. Stealing paintings done by Vermeer*
> *D. Telling the police a series of lies*

This required candidates to find the correct part of the passage, (identified as the 'target text'), namely:

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

They would then need to use *lexical knowledge* to recognise the link between 'accused' in the text item and 'arrested' in the text, *establish the propositional meaning* of the second sentence, and also *integrate information across the two sentences* in the text, as well as distinguishing between the correct answer and the distractors.

In terms of eye tracking results, when the Mann-Whitney U test was applied to test differences between successful and unsuccessful candidates' gaze patterns on the key Areas of Interest, no significant difference was noted between successful and unsuccessful candidates in terms of their attention to the target sentence or to the test item (correct and incorrect options), with one exception. This was in terms of Visit Duration on the target text sentence, where it was clear that of the candidates whose gaze data was of acceptable quality for analysis for this item, the successful students spent significantly longer on each visit to the target text than unsuccessful candidates (see *Table 6*).

Table 6. Statistics for Item 13 Target Text: Visit Duration

	N	Median	Mean	SD	Max	Min	Mann-Whitney U	Wilcoxon W	Z	p (2-tailed) p<0.05
Successful	10	1.05	1.21	0.66	2.45	0.40	8.00	23.00	-2.08	**.037**
Unsuccessful	5	0.50	0.55	0.18	0.85	0.37				
Total	15	0.79	0.99	0.63	2.45	0.37				

Once again caution should be used in drawing firm conclusions from this, given the small sample, but it would appear to support the conclusion drawn in Bax (2013b) that successful candidates are better able to find and identify the appropriate part of a text, using strategic skimming and scanning techniques, and that they can then focus on it productively, while unsuccessful candidates who are less strategic fail to do so.

Item 14

(High-Intermediate Level, Reading comprehension, Article with MCQs)

As noted in the previous section, the second item we identified from this section of the High-Intermediate Level test required understanding of the whole text (item14). This was an important item in our analysis, since it tests the higher order cognitive ability identified in Khalifa and Weir's scheme (see *Table 1*, page 1), of *"Creating a text level structure"*. This was the item itself:

What does the article indicate about van Meegeren?

A. His reputation as an artist surpassed Vermeer's.

B. He showed a painting by Vermeer in court.

C. He was admired for his artistic talent.

D. The charge against him was reduced. (correct answer)

Clearly, although the correct answer is to be found in the last paragraph, the distractors oblige

candidates to read the whole text so as to eliminate incorrect possibilities, as well as to use their lexical and syntactic knowledge. For this reason the Areas of Interest selected for analysis with this item included the text as a whole, as well as the final paragraph, the target sentence itself, and the test item (distinguishing between correct and incorrect answers). The key paragraph and target sentence (underlined here, but not in the original) were:

> *At his trial, van Meegeren confessed that the 'masterpiece' in question was a fake. No one, however, believed him. Van Meegeren finally convinced the judge by painting another fake Vermeer. Consequently, van Meegeren was convicted on a lesser offense - forging an artist's signature - and sentenced to a year in prison. His case fascinated the public and revealed how easily even experts can be deceived by fakes that bear famous names.*

Results from the eye tracking data showed no significant differences between successful and unsuccessful candidates in terms of attention paid to the test item itself. However, they did show a significant difference in terms of attention paid to one page of the text (the first one of two in the onscreen version), and also to the *target paragraph* on the second page, and (separately) to the *target sentence*. In all three cases, when the Mann-Whitney U test was applied to the data from candidates whose gaze data was of acceptable quality, significant differences were noted between successful and unsuccessful candidates. Successful students spent significantly longer on these three target areas than unsuccessful candidates (see *Table 7*). Specifically, successful candidates showed more fixations on page 1 (*Fixation Count Mean* - implying closer reading), and on each visit which they made to both the target paragraph and sentence they spent significantly longer than unsuccessful candidates (*Visit Duration Mean*).

Table 7. Statistics for Item 14

Text page 1 ***Fixation Count Mean***

	N	Median	Mean	SD	Max	Min	Mann-Whitney U	Wilcoxon W	Z	p (2-tailed) p<0.05
Successful	8	269.00	283.63	89.53	505.00	202.00	15.00	60.00	-2.02	**.043**
Unsuccessful	9	197.00	207.11	80.07	352.00	34.00				
Total	17	240.00	243.11	92.87	505.00	34.00				

Target paragraph ***Visit Duration Mean***

	N	Median	Mean	SD	Max	Min	Mann-Whitney U	Wilcoxon W	Z	p (2-tailed) p<0.05
Successful	8	8.13	6.62	2.79	9.85	2.54	11.0	56.0	-2.406	**.016**
Unsuccessful	9	2.69	3.16	2.15	8.25	0.07				
Total	17	3.02	4.79	3.05	9.85	0.07				

Target sentence ***Visit Duration Mean***

	N	Median	Mean	SD	Max	Min	Mann-Whitney U	Wilcoxon W	Z	p (2-tailed) p<0.05
Successful	8	1.67	2.07	0.98	3.84	0.62	10.00	46.00	-2.312	**.021**
Unsuccessful	8	0.91	1.00	0.48	1.46	0.16				
Total	16	1.47	1.53	0.94	3.84	0.16				

It is not methodologically sound to attribute too great a significance to these results, given the

small sample, but they are nonetheless indicative of the phenomenon noted also in Bax (2013b), namely that successful candidates tend a) to read more intently, and b) to identify and then pay significantly greater attention to key parts of the text (as was also seen with Item 13 above).

In terms of the test analysis, this suggests that this key item, which aimed to test higher order cognitive processes in Khalifa and Weir's terms, was indeed successful in distinguishing between candidates who could identify and extract key information from longer texts, and those who could not, by eliciting from them the necessary higher order cognitive processing activity. In other words, the item succeeded in eliciting from candidates higher order processing activities in their reading, and then rewarding those who succeeded in doing so.

Advanced Items

Item 19

(Advanced Level, Careful reading, Summary – fill in blanks)

In the Advanced section of the test, Item 19 was identified for analysis as it was a good representative of the items as a whole and was answered successfully by approximately half the candidates (10 out of 24). In keeping with the advanced aims of the test, candidates were required to read a text and then complete a cloze passage, the relevant section of which for this item was this:

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

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

The eye tracking gaze data, when analysed, did not reveal significant differences between the successful and unsuccessful candidates on this item in terms of focus on or visits to the cloze text itself, but it did reveal a significant difference in terms of one measure, namely Visit Count, in relation to the specific Target Phrasing (i.e. the precise short passage in the text itself which candidates had to identify and then read intensively). The statistics can be found in *Table 8.*

Table 8. Statistics for Item 19

Visit Count Target Phrasing Sum (count)

	N	Median	Mean	SD	Max	Min	Mann-Whitney U	Wilcoxon W	Z	p (2-tailed) p<0.05
Successful	5	12.00	10.90	4.18	15.00	3.00	11.00	89.00	-2.01	**.044**
Unsuccessful	12	6.50	6.25	5.34	22.00	1.00				
Total	17	7.00	7.62	5.46	22.00	1.00				

This refers to the number of *visits* which the candidates made to that area of text, and not (as with previous items) to the number of eye fixations. This implies that successful candidates had identified that this was a key area to which they had to attend, and then came back to it repeatedly, significantly more than unsuccessful candidates did. This is illustrated in **Figure 4**, where it can be seen that one successful candidate has paid far more attention to the target phrasing than the unsuccessful candidate. (The GazePlot of course includes more information, besides that of Visits – for example it shows the whole paragraph and not only the precise phrasing which was the focus of the data in *Table 8* – but it can still serve to illustrate the relative attention paid to that key text segment.)

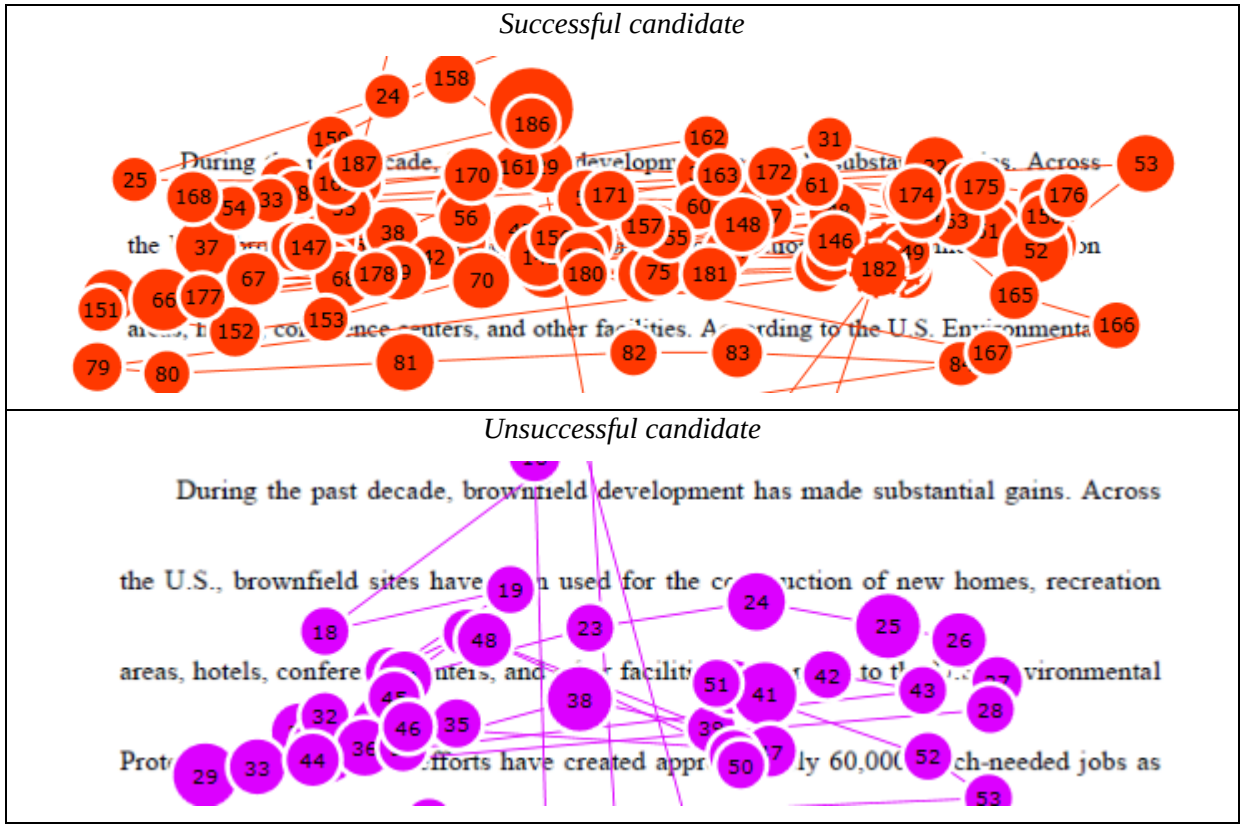


Figure 4. Gaze data for Target Text for Item 19

This is precisely the kind of cognitive activity which we would expect an advanced level test

item to elicit, since it implies that good candidates need to look at the cloze section and then back at the target section of the text a number of times in order to apply their lexical and other knowledge. This data shows that, with this item, the successful candidates did precisely that, although with these small samples such an analysis must remain suggestive rather than conclusive.

Item 21
(Advanced Level, Skimming and scanning, Headings matching)

This item required candidates to read a set of paragraphs and then to match them with the correct heading from a set of 11, a cognitively demanding task requiring at least *word recognition*, *lexical access*, *integrating information across sentences* and *inferencing*, in Khalifa and Weir's terms (see *Table 1*). The five Areas of Interest which were analysed consisted of the whole text paragraph, two key items of lexis, the set of answer as a whole, and the correct answer.

There was no significant difference identified between successful and unsuccessful candidates on most of these areas, with the exception of Visit Duration in the relevant paragraph itself, in other words the whole Target Paragraph which readers had to read so as to identify the correct heading. Successful candidates spent significantly longer in *each visit* they made than unsuccessful candidates, as can be seen in *Table 9*.

Table 9. Statistics for Item 21

Visit Duration on the Target Paragraph

	N	Median	Mean	SD	Max	Min	Mann-Whitney U	Wilcoxon W	Z	p (2-tailed) p<0.05
Successful	8	6.78	7.31	2.56	11.56	3.53	12.00	48.00	-2.10	**0.36**
Unsuccessful	8	4.26	5.25	3.76	14.90	2.28				
Total	16	5.33	6.28	3.38	14.90	2.28				

This indicates that successful candidates were obliged by the test item to pay close attention to the relevant paragraph, and did so productively, in comparison with unsuccessful candidates. This in turn suggests that the test item is working effectively and achieving its aims.

Item 30
(Advanced Level, Skimming and scanning)

This item required candidates to read a set of longer texts and then to answer a question which required understanding of detail in the text. In this case the question was:

> *Which historical attraction offers theatrical productions?*

The answer was in the first of the three texts, which included the lexis 'street performances', 'dramatic events' and 'acted out' to match the term 'theatrical productions' in the test question. To answer correctly would require the candidates to read all three texts, and use at least the cognitive processes identified by Khalifa and Weir as *word recognition, lexical access, integrating information across sentences* and *inferencing.* In addition, they would need to some extent to use the most complex of processes, namely *integrating information across texts* (see *Table 1*) since they would need to contrast the information in each of the longer texts in order to rule out possible wrong answers. Four Areas of Interest were analysed, focussing respectively on the set of questions, the specific Test question for Q30, one key phrase in the text with key pieces of lexis, and the whole of the correct page of text.

In the event the analysis identified no significant differences between successful and unsuccessful candidates in terms of their reading of the test questions or the target phrase with key lexis, but it did identify a significant difference in Visit Count for the relevant page on which the answer was to be found (see *Table 10*).

Table 10. Statistics for Item 30

Visit Count to whole text

	N	Median	Mean	SD	Max	Min	Mann-Whitney U	Wilcoxon W	Z	p (2-tailed) p<0.05
Successful	9	12.00	13.78	7.07	29.00	3.00	12.50	40.50	-2.03	**.043**
Unsuccessful	7	6.00	8.00	2.67	12.00	5.00				
Total	16	11.00	11.25	6.28	29.00	3.00				

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

4.3 Summary of eye tracking data

The data gathered from the eye tracking technology was illuminating in several ways. It must be reiterated once again that the samples are small, owing to the fact that data from each candidate has to be gathered individually, in a highly time-consuming process, and for this reason the results presented in this section must be seen as indicative as opposed to conclusive.

Although each test item analysed above produced different measures, much of the data pointed in the same direction, namely that successful candidates on each test item were probably successful because they focussed their gaze either for *longer*, or *more frequently*, or *more productively* on key areas of the test items or the texts. This in turn suggests that they were better able to identify those areas on which it is most strategic to focus. Eye tracking

cannot give direct insights into important areas of text processing, such as lexical knowledge, syntactic decoding ability, or memory, and this must be taken into account when considering the results discussed here, but nonetheless the research was successful in indicating – in line with previous literature – that the gaze behaviour of successful L2 candidates in test conditions does differ significantly at key points from the gaze behaviour of unsuccessful candidates, in ways which can help us to understand how better to train readers and how better to test them.

4.4 Implications for test design

A further important finding from the eye tracking data discussed above is that the GEPT test appears to be functioning effectively in terms of cognitive processing in the areas analysed. It is clear from all of the test items discussed above that they are successfully leading candidates to carry out lower level cognitive operations (e.g. Item 3 and 9) and also the higher level reading activities (e.g. Items 13, 14, 19, 21 and 30) requiring the type of high level cognitive processing identified by Khalifa and Weir and others as appropriate for testing reading at higher intermediate and advanced levels.

It is also important to note that in each case it was the successful test takers who exhibited the target cognitive processing, meaning that the test items are effective in distinguishing them from those test takers who do not employ the relevant cognitive processes. This is an important new piece of evidence in building a validity argument for the GEPT tests under investigation.

4.5 Self-report processing

An important caveat when dealing with eye tracking data is the fact that it is not always possible with confidence to interpret readers' cognitive processes using gaze data alone, for which reason it was important in this project also to collect self-reports and stimulated recall data in the way described in the methodology section above (See 3.5 Data collection, page 6). We can now turn to look at the findings from these two sources.

Table 11 presents a summary of the self-report findings, gathered from the reports completed by each participant after they had completed the test (using the Reading Processing Checklist in, on page 312 below). It will be seen that the summary distinguishes between candidates in terms of their test scores, although once again it should be noted that all the participants tested were at high levels of proficiency.

Table 11. Summary of self-report findings

			% of participants in each group choosing each option			
		Group	Low	Low-medium	Medium-High	High
			(n=0)	(n=0)	(n=15)	(n=9)
Cloze	Q1. Reading goal	A. to understand specific information and details	N/A	N/A	100.00	77.78
		B. to understand main idea of each paragraph	N/A	N/A	13.33	22.22
	Q2. Cognitive processing	A. test-taking strategies, e.g. matching key words	N/A	N/A	100.00	22.22
		B. Search reading	N/A	N/A	80.00	44.44
		C. Careful reading	N/A	N/A	6.67	55.56
		D. Inferencing	N/A	N/A	0.00	0.00
	Q3. Level of comprehension	A. Intra-sentential	N/A	N/A	13.33	77.78
		B. Inter-sentential	N/A	N/A	93.33	44.44
			(n=1)	(n=5)	(n=8)	(n=10)
Comprehension	Q1. Reading goal	A. to understand specific information and details	0.00	0.00	62.50	100.00
		B. to understand main idea of each paragraph	100.00	20.00	100.00	100.00
		C. to understand main idea of whole text	0.00	100.00	25.00	0.00
	Q2. Cognitive processing	A. test-taking strategies, e.g. matching key words	100.00	40.00	12.50	0.00
		B. Search reading	100.00	20.00	75.00	100.00
		C. Careful reading	0.00	100.00	100.00	100.00
		D. Inferencing	0.00	0.00	0.00	0.00

	Q3. Level of comprehension	A. Intra-sentential	100.00	0.00	0.00	0.00
		B. Inter-sentential	100.00	100.00	100.00	100.00
		C. Text-level	0.00	40.00	100.00	20.00
			(n=15)	(n=7)	(n=2)	(n=0)
Summary	Q1. Reading goal	A. to understand specific information and details	26.67	0.00	0.00	N/A
		B. to understand main idea of each paragraph	93.33	71.43	100.00	N/A
		C. to understand main idea of whole text	40.00	100.00	100.00	N/A
	Q2. Cognitive processing	A. test-taking strategies, e.g. matching key words	86.67	28.57	0.00	N/A
		B. Search reading	40.00	14.29	100.00	N/A
		C. Careful reading	53.33	100.00	100.00	N/A
		D. Inferencing	13.33	71.43	0.00	N/A
	Q3. Level of comprehension	A. Intra-sentential	20.00	0.00	0.00	N/A
		B. Inter-sentential	80.00	71.43	100.00	N/A
		C. Text-level	80.00	100.00	100.00	N/A
			(n=6)	(n=14)	(n=2)	(n=2)
Skimming & Scanning	Q1. Reading goal	A. to understand specific information and details	66.67	71.43	0.00	0.00
		B. to understand main idea of each paragraph	66.67	71.43	0.00	100.00
		C. to understand main idea of whole text	16.67	0.00	100.00	100.00
	Q2. Cognitive processing	A. test-taking strategies, e.g. matching key words	66.67	28.57	0.00	0.00

		B. Search reading	83.33	100.00	100.00	100.00
		C. Careful reading	50.00	50.00	100.00	0.00
		D. Inferencing	0.00	0.00	0.00	100.00
	Q3. Level of comprehension	A. Intra-sentential	16.67	7.14	100.00	0.00
		B. Inter-sentential	66.67	92.86	100.00	50.00
		C. Text-level	66.67	50.00	100.00	100.00
		D. Inter-textual	16.67	7.14	0.00	100.00

A sample of participants were then asked to complete a stimulated recall interview in the manner described above in section *3.5 Data collection* on page 6 above. A central aim of this was to gain more information about their cognitive activities, to add to what they and other candidates had reported in the Reading Processing Checklist as set out in summary in *Table 11* above. For this reason, both sources of data will now be considered together as we summarise what candidates said about their own processing while completing the test.

4.6 Stimulated recall data findings

Given that gaze data alone cannot tell us what a candidate's cognitive operations were at any one time, the Stimulated Recall element of our procedure was of particular value. As reported above in *3.5 Data collection* (page 6), a sample of participants were asked immediately after the eye tracking process to watch their own gaze behaviour and at the same time report on what they were doing and why. Given potential limitations in post hoc questionnaires, where participants are in danger of reporting falsely or partially, owing to the gap between the experience and the self-report, this immediate viewing of, and commenting on, their own reading was felt to be a potentially superior means of gaining insights into candidates' cognitive processes. It therefore offers us important answers to our three Research Questions, as will be apparent below.

Appendix 2 (page 33) presents key samples of interview data typical of all responses, divided into groups according to candidates' performance on the test itself, so that the first column shows typical interview comments from the low-scoring group, and so on. The data can now be discussed in detail, following the four different sections of the test as set out in *Table 2* (reproduced again here for convenience).

Table 12. Table 2 reproduced

Level	Items in GEPT test	Test item numbers in the current research project	Items identified for detailed analysis	Cognitive processes involved (minimum) *(see Table 1, page 1)*
High-Intermediate Level (14 mins)	Part 2 Cloze (n=7) • Q16-22 (MCQ)	1-7	3	-Syntactic parsing
	Part 3 Reading Comprehension (n=7) • Q37-38: Graph (MCQ) • Q46-50: Article (MCQ)	8-9 10-14	9 13 14	-Inferencing -Establishing propositional meaning at clause and sentence level -Integrating information across sentences -Creating a text level structure
Advanced Level (25 mins)	Part 1 Careful reading (15 mins) (n=6) • Q15-20 (Summary - Fill in the blanks)	15-20	19	-Word recognition -Lexical access -Integrating information across sentences -Inferencing -Creating a text level structure
	Part 2 Skimming & Scanning (10 mins) (n=10) • Q21-26 (Headings matching) • Q33-36 (Which text)	21-26 27-30	21 30	-Word recognition -Lexical access -Integrating information across sentences -Inferencing -Integrating information across texts
Total: 39 minutes	**30 test items**	**30 test items**	**7 items**	

Starting with Section 1, the ***Cloze*** section, this was analysed initially as testing syntactic parsing for the most part. It was clear from the interview data that students themselves saw the centrality of grammar knowledge in this section, with one of the high-achieving students, for example, remarking:

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

However, several of them noted that this section also required higher order cognitive skills, since it called also for inter-sentential analysis. As two of them said:

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

This reference to inter-sentential activity is important corroboration of the fact that this Cloze task is relatively high in level, as is appropriate for a High-Intermediate test, since in Khalifa and Weir's terms it is calling on higher order cognitive processes as well as lower ones.

Section 2, ***Comprehension***, begins with a task involving reading a graph and answering some MC questions. Scores were high on this section, and indeed candidates reported that they found it relatively straightforward, one stating that:

This section was quite easy. I read the diagrams first to get some information out of them. I then read the questions and focused on each option to see which one was correct.

A high-scoring candidate reported that:

I read the diagram to understand what it was about and then read the questions. I then read the diagram again to find the specific information, e.g. figures of different countries.

In terms of cognitive processing, this part was initially analysed as expecting *inferencing*, and indeed this cognitive process was reported by candidates, for example the one who said that *"I noticed the information for different quarters"*. Candidates also reported on the need to use search reading – in this case reading of the diagram.

The second part of this section was a longer article which called for a range of more advanced cognitive processes. A good description of these processes and how they interconnected was offered by a high-scoring participant:

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

This illustrates also the element of search reading used in this part of the test, and also the need for *Establishing propositional meaning at clause and sentence level*, of *Integrating information across sentences* and also of *Creating a text level structure* for some of the test items. In short, the interview data shows that all of these cognitive processes were used by participants as expected in this section.

Section 3, the first of the Advanced parts of the test, was the ***Summary***. It will be recalled that student performance on this section was the weakest (see *4.1 Performance* on page 8 above), and in the stimulated recall interviews candidates duly reported that it was the most difficult. The summary here did not merely make use of words copied from the original text; as one candidate noted:

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

This added a degree of cognitive complexity, with an additional lexical and syntactic parsing load. Indeed many weaker candidates reported that they fell back on relatively basic cognitive strategies by trying to match words in the summary and the passage. For example these two low-scoring candidates reported as follows:

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

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

The last comment from the second speaker indicates awareness that this strategy was not very effective. One candidate tried to use not only lexical but meta-discourse devices to assist:

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

It is clear from these last two examples that such lower level cognitive strategies were seen to be inadequate with these test items. By contrast, more successful candidates specifically reported using higher order cognitive processes such as inferencing, as in these two instances:

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

…sometimes I tried to understand why the author mentioned particular details, for example different countries, in a particular place. And for example I was thinking what the similarities or differences did these countries have.

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

The final section, also at Advanced level, was ***Skimming and Scanning***. Candidates at all levels of performance reported carrying out the kinds of cognitive activity expected of these items, but at varying degrees of sophistication. To start with, this lower scoring candidate reports a relatively basic cognitive operation:

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

By contrast, this higher scoring candidate reports rather more sophistication and detail:

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

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

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

Furthermore, it is clear from this that the candidate's report also made use of high order inter-sentential reading, as well as search reading.

4.6.1 Discussion of self report and stimulated recall data

A. Test-taking strategies

If we turn now to consider the implications of the self report and stimulated recall interview data, it is clear from the information in *Table 11* above, and also from interview data discussed above and exemplified more fully in *Appendix 2* (page 33), that lower-scoring participants tended to report the use of more *test-taking strategies*, e.g. matching key words, focusing on topic sentences, getting hints from grammar, eliminating options, etc. This can be seen for example in the *Cloze section*, *Q2 Cognitive processing*, and also in the same section of the *Comprehension* section, where 100% of the lower scoring candidates reported using such strategies. In later sections too, lower candidates reported similar use of such test-taking strategies.

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

One possible reason for this, alluded to by some participants, was a negative washback effect resulting from their extensive test preparation at school. Candidates reported that they had at school received extensive tuition in how to complete reading tests, and it appeared as if some of them were unable to escape from this 'conditioning' even though they knew that with more

advanced test items, calling for higher order cognitive skills, they might be better not to fall back on such prepared strategies and 'tricks'.

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

B. Range of cognitive processes

A second finding from the self report and interview data, as can also be seen in *Table 11*, is that candidates employed the full range of cognitive processing which we would expect of a high-intermediate and advanced level test. This will be considered again when we revisit the Research Questions below.

5. CONCLUSIONS AND RECOMMENDATIONS

5.1 Research Questions revisited

The first Research Question was as follows:

1. What cognitive processes are elicited by different sections of the GEPT High-Intermediate Level Reading Test?

The eye tracking data alone could not inform us with certainty as to the cognitive processes which a reader was employing at any moment, but it did give strong indications which were then corroborated through the stimulated recall interviews and processing questionnaires. In addition, it is legitimate to infer that if a test item requires a certain cognitive process from a candidate (e.g. inferencing), and the candidate answers the item correctly, then also offers gaze date to show the appropriate eye movements (e.g. focusing on the target sentence), and then also reports having read the text in that way, we can reasonably infer that the target cognitive process has been used.

Taking into account all these sources of data, namely the item analysis summarised in *Table 2* (page 5), performance data for each candidate, together with the analysis of actual cognitive processes as implied by the eye tracking data for the *High-Intermediate* items analysed (pages 11-15), together also with the retrospective reports and stimulated interviews reported above, it is clear that the High-Intermediate section of the GEPT test successfully elicited and tested the following lower and higher cognitive processes:

-Word recognition
-Lexical access
-Syntactic parsing
-Inferencing
-Establishing propositional meaning at clause and sentence level

-Integrating information across sentences
-Creating a text level structure

Research Question 2 was as follows:

2. What cognitive processes are employed by test-takers on different sections of the GEPT Advanced Level Reading Test?

Using the same data sources, it was possible also to conclude that the Advanced sections of the test elicited the same set of cognitive processes as the High-Intermediate test, with the addition in the final section of the most difficult of all in Khalifa and Weir's scheme, namely:
-Integrating information across texts.

However, this last element was only tested partially, since candidates in that section needed to read across different texts only so as to exclude possible wrong answers, and not precisely to "integrate information" in any complex way. If the test designers wished to raise the cognitive level of this Advanced test, then a more fully 'intertextual' activity could arguably be added to the GEPT Advanced paper, by which candidates could be asked to demonstrate that they had read and assimilated information across more than one complex text.

However, in summary, it was apparent from our data that the Advanced level test items were indeed testing the higher order cognitive skills in Khalifa and Weir's scheme, as they were designed to do.

Finally, Research Question 3 was as follows:

3. To what extent and in what ways do the cognitive processes elicited at the two levels match the cognitive processes anticipated in reading tests at these levels?

Returning to the scheme set out by Khalifa and Weir, summarised above in *Table 1* on page 1, as well as to their larger discussion of what a High-Intermediate and Advanced reading test should require of candidates in terms of cognitive processing, it is apparent that the two elements of the GEPT test which were investigated in this project succeed in requiring of candidates the range of cognitive processing activity commensurate with High-Intermediate and Advanced reading levels respectively. Our results show that candidates who answered correctly were using notably different gaze patterns in some cases, and furthermore that they were then able to report retrospectively that they had used higher order cognitive processes as they were reading and responding.

Finally, the project was also notably successful in combining the innovative use of advanced eye tracking technology with more traditional paper questionnaires, and with innovative stimulated recall procedures. By these means it was possible to offer unprecedented and illuminating insights into readers' behaviour when completing high level reading test items.

References

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

Bax, S. (2012). Cognitive processing of candidates during reading tests: Summary evidence from two eye-tracking projects. *Selected Papers from the 21st International Symposium on English Teaching*, Taipei, Taiwan.

Bax, S. (2013a). Readers' cognitive processes during IELTS reading tests: evidence from eye tracking. *ELT Research Papers,* 13-36. Retrieved from http://www.teachingenglish.org.uk/article/readers%E2%80%99-cognitive-processes-during-ielts-reading-tests-evidence-eye-tracking

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

Baxter, G. & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practices, 17*(3), 37-45.

Cohen, A. D. & Upton, T. A. (2007). 'I want to go back to the text': Response strategies on the reading subtest of the new TOEFL. *Language Testing, 24*(2), 209-50.

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

Eger, N., L. Ball, R. Stevens & Dodd, J. (2007). Cueing retrospective verbal reports in usability testing through eye-movement replay. *Proceedings of HCI 2007, The 21st British HCI Group Annual Conference University of Lancaster, UK. http://www.bcs.org/server.php?show=ConWebDoc.13300*

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

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

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

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

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

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

Tobii Technology AB (2013). *Product Description Version 1.0.1,* Retrieved from http://www.tobii.com/Global/Analysis/Downloads/Product_Descriptions/Tobii_X2_Product_Description.pdf?epslanguage=en

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

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

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

Appendices

Appendix 1: Reading Processing Checklist

	Part 1 (Q1-7)	**Part 2** (Q8-14)	**Part 3** (Q15-20)	**Part 4** (Q21-30)
Q1. To find the answer to the questions in this part, I tried to read for ... *(Please choose 1 option only)*				
a) **specific information** and **details**;	a	a	a	a
b) main idea of **each paragraph**; or	b	b	b	b
c) main idea of **whole text**.		c	c	c
Q2. I found the answer to the questions in this part mainly by ... (Y*ou can choose more than 1 option*)				
a) using test taking-strategies, such as quickly **matching words** in the question with similar words;	a	a	a	a
b) **searching** quickly for **part(s)** of the text which might answer the question;	b	b	b	b
c) reading the **whole** text **slowly** and **carefully** to find the answer to the question; and/or	c	c	c	c
d) **guessing** ideas which are not explicitly stated.	d	d	d	d
Q3. I found the answer to the questions in this part mainly ... (Y*ou can choose more than 1 option*)				
a) within **a single sentence**;	a	a	a	a
b) across **sentences** within a paragraph;	b	b	b	b
c) across **paragraphs** within a text; or		c	c	c
d) across **texts**				d

Appendix 2: Samples of typical comments by candidates at different levels during the stimulated retrospective protocol

Group			Typical comments from participants in each performance group			
			Low (n=0)	Low-medium (n=0)	Medium-High (n=15)	High (n=9)
Section 1 Cloze	Q1. Reading goal	A. to understand specific information and details				
		B. to understand main idea of each paragraph				
	Q2. Cognitive processing	A. test taking strategies, e.g. matching key words			-first I read through the questions and tried to find out what options would fit the blank best. I pay attention to the sentence before and after the blank.	-I focused on my grammar knowledge when deciding the answer and to check if the answer makes the sentence coherent
		B. Search reading			-I did this section quite fast	
		C. Careful reading				-I read the whole cloze sentence by sentence to understand the overall meaning. - read the cloze sentence by sentence to have a rough idea. I actually tried to fill in the blanks by myself first before looking at the options. I then read the options to choose the best one.
		D. Inferencing				
	Q3. Level of comprehend-sion	A. Intra-sentential				
		B. Inter-sentential			-…I believe in this section you have to understand the meaning of every sentence, and the relation between sentences to fill in the blank.	

			(n=1)	(n=5)	(n=8)	(n=10)
Section 2 **Comprehension**	Q1. Reading goal	A. to understand specific information and details				
		B. to understand main idea of each paragraph				
		C. to understand main idea of whole text				
	Q2. Cognitive processing	A. test-taking strategies, e.g. matching key words		-This section was quite easy. I read the diagrams first to get some information out of them. I then read the questions and focused on each option to see which one was correct.		
		B. Search reading		-I scanned the questions and started looking for answers and then read about the options.	-I got the overall topic of the diagram from the title. And I noticed the information for different quarters. It wasn't difficult to locate the specific details. I got the overall meaning of the whole article by reading the first and last sentence of each paragraph quickly. I then read the questions and located the relevant parts in the article for each question. The article was structured in sequence so it wasn't too difficult to locate the relevant parts.	-I read the diagram to understand what it was about and then read the questions. I then read the diagram again to find the specific information, e.g. figures of different countries. -I read the chart first to identify the main idea and to understand the different sections of the chart. And then I read the questions and then went back to find the specific details.
		C. Careful reading				
		D. Inferencing				
	Q3. Level of comprehension	A. Intra-sentential				
		B. Inter-sentential				
		C. Text-level		-I needed to understand the overall meaning of the passage before I could do anything.	-I got the overall meaning of the whole article by reading the first and last sentence of each paragraph quickly.	

			(n=15)	(n=7)	(n=2)	(n=0)
Section 3 **Summary**	Q1. Reading goal	A. to understand specific information and details				
		B. to understand main idea of each paragraph				
		C. to understand main idea of whole text				
	Q2. Cognitive processing	A. test-taking strategies, e.g. matching key words	-It was the most difficult section. I did not understand most of the words. The summary has been paraphrased and most of the words have been changed from the original passage. To find the answer, I tried to focus on the connectives, e.g. however, to guide me to the relevant part. But it was difficult I couldn't find most of the answers. I tried to match some key words between the summary and the original text, e.g. authority and the UK government -This section was really difficult. I couldn't see the connection between the summary and the original article. I was then trying to identify some key words in the summary which link back to the article. I was lost. -I was then looking for some words in the summary which relate to the original article. - I also use some hint from the blank e.g. what tense it should be and the part of speech, to help finding the answer. If it should be a verb, I focused on finding a verb from the article. -I tried to identify some relevant vocabulary.			
		B. Search reading				
		C. Careful reading				
		D. Inferencing		-Sometimes I made a guess based on my understanding. For example, it was talking about		

				industries closing down so it means they moved out of the area. - And sometimes I tried to understand why the author mentioned particular details, for example different countries, in a particular place. And for example I was thinking what the similarities or differences did these countries have.		
	Q3. Level of comprehension	A. Intra-sentential				
		B. Inter-sentential				
		C. Text-level				

			(n=6)	(n=14)	(n=2)	(n=2)
Section 4 **Skimming & Scanning**	Q1. Reading goal	A. to understand specific information and details				
		B. to understand main idea of each paragraph				
		C. to understand main idea of whole text				
	Q2. Cognitive processing	A. test-taking strategies, e.g. matching key words	-My main strategy was to identify the key words in each option like 'historical' and then looked for the relevant paragraph in the article. But it was difficult because all paragraphs were sort of related. I think this strategy was wrong because I only had 10 minutes for this section. And time was running out. -I just skim the paragraph and choose my answer based on some keywords I noticed. -I focused on some keywords, like motorway, which may help me to find the heading.	-I read all the paragraphs very quickly first. And then I read the options of the heading. I then read the paragraphs again and was trying to find some key words which are relevant to each heading or to match the exact words from the heading. -I first skimmed the headings to identify some keywords to get me a general idea and then checked if I could see these keywords or similar words in every paragraph. I went back and forth to see if I could match the keywords. I also answered those questions which I thought were easier first.		
		B. Search reading	-After reading each paragraph, I scanned every heading to see which one was most suitable.		-I had time pressure in this section, so I read the paragraphs very quickly. I needed to understand the overall meaning of each paragraph before choosing the most suitable heading. Some headings were quite similar so I had to understand the meaning of the paragraph. If I only relied on some key words, they would probably misguide me.	
		C. Careful reading				
		D. Inferencing				

	Q3. Level of comprehension	A. Intra-sentential				
		B. Inter-sentential	-I tried to find out the meaning from the first few sentences. I then realised this wasn't useful I couldn't get much information about which heading was most suitable. So I had to read the whole paragraph one by one.		-It was like reading the options getting the main idea of each one, and then read the paragraphs to see which heading fits which paragraph best.	
		C. Text-level				
		D. Inter-textual				

The Language Training and Testing Center (LTTC)
No.170, Sec.2, Xinhai Rd., Daan Dist.,
Taipei City, 10663 Taiwan(R.O.C.)
Tel: +886-2-2377-8071
Email: geptgrants@lttc.ntu.edu.tw
Website: www.lttc.ntu.edu.tw