UNIVERSITY of CAMBRIDGE
ESOL Examinations

# Research Notes

## Issue 47

February 2012

ISSN 1756-509X

UNIVERSITY *of* CAMBRIDGE
ESOL Examinations

# Research Notes

Issue 47 / February 2012
A quarterly publication reporting on research, test development and validation

Printed in the United Kingdom by Océ (UK) Ltd.

# Research **Notes**

## Contents

## Editorial notes

Welcome to issue 47 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within University of Cambridge ESOL Examinations.

This issue – the first of 2012 – presents the research outcomes from the first round of Cambridge ESOL's Funded Research Programme undertaken in 2010. It benefits from the guest editorship of Dr Jayanti Banerjee, Program Manager at Cambridge Michigan Language Assessments.

Following Dr Banerjee's guest editorial (see the following page) which describes the projects and suggests their impact for Cambridge ESOL and more widely, there are four articles based on the Cambridge ESOL Funded Research Programme which cover a range of topics and contexts relevant to the teaching or testing of Cambridge English. The reported research includes investigations of the validity of test items and candidates' output, and the impact and use of various Cambridge English tests in two specific contexts. Such studies enable Cambridge ESOL to support research that goes beyond the normal range of studies we are able to commission or undertake ourselves, thereby enhancing our understanding of the nature and impact of the language tests we work with on a daily basis, and additionally providing important outsider viewpoints from both established and newer researchers in the language testing – or teaching – fields.

The second round of research funded by this programme is close to completion, and the third round is already underway, so we look forward to reporting on these studies in future issues of *Research Notes*. For those readers inspired to submit their own research proposals, the Call for Proposals for the fourth round is expected to be available in August 2012 on the Cambridge ESOL Research and Validation website, so for further details visit www.research.CambridgeESOL.org later this year.

We finish this issue with an update on ALTE events from Martin Nuttall of the ALTE Secretariat; the announcement of the winners of the Caroline Clapham IELTS Masters Award 2011 and the 2012 Cambridge/ILTA Lifetime Achievement Award, and details of the 30th volume to be published in the *Studies in Language Testing* series.

With the new calendar year we are thinking of introducing various innovations to *Research Notes*, and are planning a reader survey later this year to help inform the future direction of this publication.

# Guest editorial

**JAYANTI BANERJEE** CAMBRIDGE MICHIGAN LANGUAGE ASSESSMENTS, USA

English language tests matter. They matter for the children who are compiling their language portfolios as well as for young adults hoping to study in an English-medium university. They matter for university admissions personnel or employers who are selecting the best candidates for their degree programmes or jobs. English language tests have tremendous symbolic power (Shohamy 2001:118) because they confer access to privileges, certify, and by extension, delimit knowledge.

As a result, providers of English language tests have a great responsibility to stakeholders. Test users rely on test developers to provide high-quality tests that meet professional standards. They also expect testing organisations to present evidence to support test score interpretations and uses. Cambridge ESOL takes these professional responsibilities seriously and has developed a Principles of Good Practice booklet (www.CambridgeESOL.org/about/standards/pogp.html) that encapsulates the organisation's commitment to five essential principles: validity, reliability, impact, practicality and quality.

As part of this commitment, in late 2009 the organisation launched the Cambridge ESOL Funded Research Programme. The first Call for Proposals encouraged studies of its Cambridge English exams in the following areas:

- test validation issues
- issues relating to contexts of test use
- issues of test impact.

This issue of *Research Notes* showcases the four projects that were funded in the first round and which took place in 2010. Each study provides insight into one or more Cambridge English examinations in a specific context or from a specific perspective.[1]

Bax and Weir (this issue) have investigated the cognitive processes employed by participants on a computer-based *Cambridge English: Advanced (CAE)* Reading test in order to check the extent to which the items elicit the range and level of cognitive processes expected of an advanced level Reading test which seeks to emulate real-world academic reading processes. They used eye-tracking technology to collect data in the form of Gaze Plots and Heat Maps which indicate both how the volunteer test takers' eyes moved when reading the input texts and answering the questions as well as how long the test takers looked at particular sections of the text. Bax and Weir also administered questionnaires to capture immediate retrospections from test takers. The resulting data confirmed that the test takers employed an appropriate range and level of cognitive processes as targeted by *CAE* items. The paper not only provides evidence for the validity of the *CAE* Reading section but it also demonstrates the value of eye-tracking technology in test validation.

Littlemore, Krennmayr, Turner and Turner (this issue)have analysed a subset of exam scripts from the Cambridge Learner Corpus to investigate the features of metaphor that distinguish performances at different levels of the Common European Framework of Reference (CEFR, Council of Europe 2001). Using the Metaphor Identification Procedure (MIP) developed by the Pragglejaz Group (2007), Littlemore et al found that metaphor use increases with proficiency level. Metaphor clusters emerge only at the intermediate levels. Littlemore et al also found that the types of metaphors used changes with proficiency level, as well as the functions these metaphors perform. These findings suggest that descriptors for metaphor use could feasibly be incorporated into rating scales for writing.

Nagao Tadaki, Takeda and Wicking and Tsagari (this issue) have focused on test use in specific contexts. Nagao et al have investigated attitudes towards the *Cambridge English: Preliminary (PET)* in Japan, an emerging market for the test. This study is particularly interesting because the *PET* is relatively new in Japan and the study has captured knowledge about the exam as well as attitudes towards it at a very early stage of its introduction. The study shows that the test does meet learners' needs but is less popular with teachers. It identifies the need for teacher support programmes and it also sheds some light on the *PET*'s fitness for purpose in the Japanese context.

Tsagari has studied *Cambridge English: First (FCE)* test preparation classes in Cyprus. Through a combination of classroom observations and teacher interviews, Tsagari amassed a rich description of the learning activities and teacher talk. She found considerable influence of the test upon the learning activities in the classroom and also in the teacher talk, particularly the advice that teachers gave to their students. Some of this influence was very positive but there were also barriers to positive impact. Tsagari points out that the teachers were not an open conduit of information about the exam. Rather, the impact of the *FCE* upon the classroom was mediated through the teachers' knowledge and beliefs about the exam, their professional skills, and their own language ability. As such, in addition to providing a window into *FCE* preparation classes, this study has identified stakeholder needs in Greece.

Anastasi (1986:4) and Cronbach (1988) remind us that the process of gathering validity evidence is never complete. Indeed, the more important and influential a test, the greater the need for collecting ongoing evidence for the validity of its use. Together, these papers contribute to the growing body of validity evidence for the Cambridge ESOL General English examinations.

[1] For promotional purposes, Cambridge ESOL increasing refers to its exams by titles such as *Cambridge English*: *Key*; *Preliminary*; *First*; *Advanced*; and *Proficiency*, although the names of the exams themselves have not changed. Our authors frequently refer to the exams by their acronyms – *KET*, *PET*, *FCE*, *CAE*, and *CPE*, respectively. For more information, see www.cambridgeesol.org/exams

## References

Anastasi, A (1986) Evolving concepts of test validation, *Annual Review of Psychology* 37, 1–15.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment*, Cambridge: Cambridge University Press.

Cronbach, L J (1988) Five perspectives on the validity argument, in Wainer, H and Braun, H I (Eds) *Test Validity*, Hillsdale, NJ: Lawrence Erlbaum Associates, 3–18.

Pragglejaz Group (2007) MIP: A method for identifying metaphorically used words in discourse, *Metaphor and Symbol* 22 (1), 1–39.

Shohamy, E (2001) *The Power of Tests*, Harlow, Essex: Pearson Education Limited.

# Investigating learners' cognitive processes during a computer-based CAE Reading test

**STEPHEN BAX** CRELLA, UNIVERSITY OF BEDFORDSHIRE, UK
**CYRIL WEIR** CRELLA, UNIVERSITY OF BEDFORDSHIRE, UK

## Introduction

This study investigates the cognitive processes employed by participants on a computer-based *CAE* Reading test, with a view to assessing the cognitive validity of the Reading test items. It takes as its starting point the cognitive processing approach with its set of cognitive processes described by Khalifa and Weir (2009 Chapter 3). In addition it draws on the methods for investigating those processes adopted in Weir, Hawkey, Green and Devi's study of academic reading in the UK (2009), and complements and extends them using onscreen recording and eye-tracking technology, as well as developing other aspects of the methodology. The central question was to what extent the test items elicited the range and level of cognitive processes expected of an advanced Reading test which seeks to emulate real-world academic reading processes.

In the event, insights from eye tracking combined with questionnaire data to provide convincing evidence that even this limited set of *CAE* test items succeeded in eliciting a wide range of appropriate cognitive processes, including those higher level reading processes necessary for real-world academic reading. Data including Gaze Plots and Heat Maps illustrating participants' eye movements indicated that test takers successfully employed an appropriate range and level of cognitive processes as targeted by the *CAE* items. In the process the project also demonstrated that eye-tracking technology, in careful combination with more traditional methods of analysis, has the potential significantly to improve our capacity to validate Reading test items in future.

## Rationale

It is axiomatic that language tests assessing the academic language proficiency of overseas students, if they are to be appropriate for university admission, should reflect the demands of the academic courses these students are aiming to follow. In addition, international examination boards have a duty to provide valid information for stakeholders and to demonstrate quality.

One aspect of such language tests which should be demonstrably valid is the extent to which they assess the cognitive processes required in academic study. For example if an advanced Reading test is to be accepted as valid by academic institutions it should demonstrably test the range and level of cognitive processes typically expected in academic study contexts, including cognitive processing at lower and higher levels. If it fails to do so – for example if it tests only a limited range of processes or only low-level cognitive processes – then it cannot claim to be an appropriate tool for assessing the academic language competence required at university level.

This is to insist on what is known as cognitive validity. Since the 1990s it has been argued that tests assessing complex cognitive constructs should establish this sort of validity (Glaser 1991, Baxter and Glaser 1998) since cognitive interpretative claims are 'not foregone conclusions, [but] need to be warranted conceptually and empirically' (Ruiz-Primo, Shavelson and Schultz 2001:100). By the same token, Weir has argued that those language tests which implicitly or explicitly claim to match real-world behaviour should also be cognitively valid (Weir 2005). In short, if a language test does not elicit from test takers the same type and level of cognitive processing as is used and expected in the real-world target situation, then it is not a valid instrument for assessing that area of linguistic behaviour. It is these issues, concerning the range and type of cognitive processing in *CAE* onscreen reading tests, which the current project sought to investigate.

Traditionally, research into readers' cognitive processes has depended heavily on retrospective or concurrent verbal reporting as a means of understanding what readers are thinking as they complete Reading test items. Recent improvements in eye-tracking technology, however, furnish additional opportunities to gain insights into readers' actual as opposed to reported behaviour, permitting significantly enhanced insights into their ongoing, second-by-second reading activity and hence a greater insight into their probable cognitive processing.

## Researching cognitive processes in academic reading

In Weir, Hawkey, Green and Devi's study (2009) of reading for academic purposes in UK universities, a number of earlier models of reading proved to be helpful, especially those that accounted for the purposeful and strategic activities of readers in an academic context and those which specified the types of reading relevant to that academic context (see Weir et al 2009 for a full description of these). As the authors note:

> in general terms, the reading types covered [in an academic context] are expeditious reading, i.e. quick, selective and efficient reading to access desired information in a text (scanning, skimming and search reading), and careful reading, i.e. processing a text thoroughly with the intention to extract complete meanings from presented material (Weir, Hawkey, Green and Devi 2009:160)

Urquhart and Weir's (1998) distinctions between global/local and careful/expeditious are of particular importance to the design of this study as they offer a taxonomy of different types of reading which are relevant to reading academic English. *Global comprehension* refers to the understanding of information beyond the sentence, including main ideas, the links between ideas in the text and the way in which these are elaborated. It involves integrating information in the text, mental model building and understanding how macro propositions in the whole text fit together. The reader in careful global reading attempts to identify the main idea(s) by reconstructing the macro-structure of a text. Logical or rhetorical relationships between ideas are represented in complexes of propositions (see Vipond 1980), often represented by the writer by means of paragraphing; global reading involves attempting to reconstruct these complexes. *Local comprehension* concerns the understanding of propositions within the sentence (individual phrases, clauses and sentences). Local comprehension involves word recognition, lexical access and syntactic parsing and establishing explicit propositional meaning at the phrase, clause and sentence level. *Careful reading* involves extracting complete meaning from a text, whether at the global or local level. As noted above, this is based on slow, careful, linear, incremental reading for comprehension. *Expeditious reading*, in contrast, involves quick, selective and efficient reading to access relevant information in a text.

Careful reading as an umbrella term encompasses processing at sentence, intersentential, text and multi-text levels. It is important that tests designed to predict the ability to read in English at university level have a range of items which extend beyond comprehension at the sentence level, i.e. they should contain a high proportion of items that test reading at the more complex stages of processing (see Khalifa and Weir 2009 for discussion of these). In academic life readers find themselves having to read and learn from a whole text as well as integrating information from various texts, especially for the preparation of assignments. Tests which focus on sentence-level processing alone are therefore not the best indicators of academic reading ability.

Typically in the past, models of reading have usually been developed with only careful reading in mind (see, for example, Hoover and Tunmer 1993, Rayner and Pollatsek 1989). However, careful reading models have little to tell us about how skilled readers cope with other expeditious reading behaviours such as skimming for gist (Rayner and Pollatsek 1989: 477–478). Carver (1992) and Khalifa and Weir (2009) suggest that the speed and efficiency of reading is important as well as comprehension. In relation to reading for university study, Weir et al (2009:162) found that in their sample of university undergraduates 'for many readers reading quickly, selectively and efficiently posed greater problems than reading carefully and efficiently'. Khalifa and Weir (2009) distinguish three types of expeditious reading skill relevant to academic study: scanning, skimming and search reading. Scanning is a form of expeditious reading that occurs at the local level. It involves reading highly selectively to find specific words, figures or phrases in a text. Skimming is generally defined (Urquhart and Weir 1998, Weir 2005) as reading quickly by sampling text to abstract the gist, general impression and/or superordinate idea: skimming relates exclusively to global reading.

Unlike skimming, search reading involves predetermined topics. The reader does not necessarily have to establish a macro-propositional structure for the whole of the text, but is rather seeking information that matches their requirements. However, unlike scanning (where exact word matches are sought) the search is not for exact word matches, but for words in the same semantic field as the desired target information. Search reading can involve both local and global-level reading. Where the desired information can be found within a single sentence the search reading would be classified as local and where information has to be constructed across sentences it would be seen as global. Search reading at the global level is the key expeditious reading skill for university students.

Khalifa and Weir's (2009) exegesis adds a further layer to this depiction by identifying the cognitive processes that underlie the types of reading relevant to the academic context and the cognitive load imposed on that processing by the various contextual parameters of the text itself (in terms of lexical and syntactic complexity, and cohesion). They argue that reading proficiency is a function of both the level of processing required by the reading task and the complexity of the reading text it is carried out on. In this study we are focusing on the nature of the processing required by reading tasks at the item level. For details of text complexity, i.e. contextual parameters in reading, the reader is referred to Khalifa and Weir (2009 Chapter 4).

Our research study investigates participants' processing of a small number of *CAE* Reading test items. Our interest is in the extent to which the items elicit the range and level of cognitive processes relevant to academic study in English. If the items only elicit cognitive processes at a lower level of complexity (word recognition, lexical access and syntactic parsing, and establishing explicit propositional meaning at the phrase, clause and sentence level), then their validity for assessing academic reading in English is in question; if, however, our relatively small sample of *CAE* Reading test items demonstrably elicit a wider range of processing, in terms of the Khalifa and Weir processing model including integration of information, building a mental model of a text, as well as text-level comprehension, this is a positive start to establishing cognitive validity evidence for the claim of *CAE* to be an appropriate test of reading in English for academic purposes. Obviously we would eventually need to look at a larger sample of *CAE* Reading items in terms of their cognitive validity and also investigate *CAE* Reading texts in terms of their comparability to those met in academic life to establish more substantial evidence of the test's construct validity.

Khalifa and Weir's (2009) model accounts for the different types of reading that readers might choose to carry out in academic life, the different levels of processing that might be activated, and the knowledge base necessary to complete an assigned reading task successfully. This model provided us with the theoretical framework on which our onscreen retrospection questionnaire was based, and also on which our analysis of the eye-tracking data was structured.

## A processing approach to investigating reading

Weir et al (2009: 162–3) review the literature on what has been called a *subskills* approach to testing reading, which is:

> based on the assumption that it is possible to target particular types of item or test task to specific types of reading so that one item might target the ability to understand the meaning of an individual word in a text and another might target the ability to extract the overall meaning of a text within a very limited time frame (skimming).

They also note that '[t]he debate over subskills centred on the ability of expert judges to arrive at a consensus about what was being tested and the essential role of the candidate was largely overlooked. The majority of studies paid surprisingly little attention to the cognitive processing required for candidates to carry out test tasks' (Weir et al 2009:63), and then cite Alderson (2000:97) who argues that:

> [t]he validity of a test relates to the interpretation of the correct responses to items, so what matters is not what the test constructors believe an item to be testing, but which responses are considered correct, and what process underlies them.

In short, understanding of the trait being measured requires an insight into the cognitive processing required for completion of the task.

## Eye tracking in the study of cognitive processes in reading

In an attempt to gain insight into readers' cognitive processes many researchers have adopted procedures in which participants report retrospectively on the linguistic process which they have engaged in. However, given the doubts sometimes expressed about the use of retrospective reporting, for example by Afflerbach and Johnston (1984) and Cordon and Day (1996), and since eye-tracking technology has improved considerably in recent years, we decided to make use of eye tracking technology in this study in order to gain better, albeit still indirect, insight into cognitive processing in combination with a retrospective questionnaire in ways to be detailed below.

The use of eye tracking in the study of reading is not new. Rayner (1998) reviews 100 years of research into reading using eye tracking of various sorts, divided into three periods before we reach what Duchowski (2002) has called the current 'fourth era' distinguished by the possibility of interactivity. Rayner highlights some of the main insights which eye tracking has offered for our understanding of reading. Firstly, when reading English, it is noted that eye fixations (when the eye dwells momentarily on a particular point) typically last about 200–250 milliseconds and the mean saccade size (i.e. when the eye moves from one point to another) is 7–9 letter spaces (Rayner 1998:375). This is of interest in the present study, particularly when identifying individual words in a text which constitute the answer to a test item. Second, eye movements are influenced by numerous textual and typographical variables, for example 'as text becomes conceptually more difficult, fixation duration increases, saccade length decreases, and the frequency of regressions [where the eye moves back rather than forwards] increases' (ibid:376), which could potentially be useful in comparing better and worse readers, although this is not a focus of the current study.

Importantly for the current project, Rayner also notes that the basic theme of his historical review, in particular of the third era from the 1970s onwards, 'is that eye movement data reflect moment-to-moment cognitive processes' (Rayner 1998:372). He expands the point as follows:

> A crucial point that has emerged recently is that eye movement measures can be used to infer moment-to-moment cognitive processes in reading . . . and that the variability in the measures reflects on-line processing. For example, there is now abundant evidence that the frequency of a fixated word influences how long readers look at the word (Rayner 1998:376).

More recent studies concur with Rayner as to the value of eye tracking for researching cognitive processes. Spivey, Richardson and Dale (2009) offer a detailed discussion of how and why eye movements can be taken to be good indicators of cognitive processes, and term them 'a window into language and cognition' (2009:225). The same metaphor is used by Salvucci and Goldberg who see eye tracking as 'a window into observers' visual and cognitive processes' (2000:71; see also Anson, Rashid Horn and Schwegler 2009). Some researchers such as de Greef, Botzer and Van Maanen (2010) take this to extremes, suggesting – to quote the title of their article – that 'Eye-Tracking = Reading the Mind', but this is arguably over-confident. It is our position that although the technology offers possibly the best available insight into cognitive processes, eye-tracking data should be treated as merely indicative of cognitive processing, rather than a true and full reflection of it.

In terms of developments in eye-tracking technology, recent advances have improved immeasurably our ability to detect what readers are looking at second by second, allowing the detailed analysis of individual differences between readers at a very high level of detail (see e.g. Bertram 2011, Buscher, Biedert, Heinesch and Dengel 2010, Eger, Ball, Stevens and Dodd 2007), as well as the analysis of highly precise fixation and saccade patterns.

## Research methodology

In this section we outline the research design, instruments used, eye-tracking software and hardware, along with the participants.

### Research instruments

One research tool used in our study was the retrospective report. As Weir et al (2009:163) note:

> [a] process-oriented approach to defining reading activity in language tests seeks an experimental method which permits comment on the actual reading process itself.

Participants in their particular study were given one part of an *IELTS* Reading test, chosen by an expert focus group so as to include a range of items requiring both explicitly stated and implicit information located across sentences, and allowing both expeditious and careful reading types, and were then asked to complete a retrospection form. Among other things, this form investigated the processes that participants engaged in while locating the correct answer to each individual item of the Reading test.

One limitation of that approach is that test takers needed to complete the full set of test items before completing the retrospective questionnaire, so that their recall of the cognitive processes they had employed was necessarily delayed. The literature on stimulated recall (e.g. Gass and Mackey 2000) emphasises the fact that the sooner after the experience the recall is elicited, the more likely it is to be accurate. For this reason our study makes use of the flexibility of computer-based delivery in order to elicit recall of the cognitive processes immediately after completion of each test item. Given the fact that their retrospection is therefore almost immediate it was anticipated that this approach would afford greater reliability in terms of participants' introspection concerning the cognitive processes they employed for each test item.

### Design

The original *CAE* test used, produced by Cambridge ESOL in Adobe Flash format, was reproduced so as to be identical in every way (using Adobe Flash) and allow maximum control over font size, interactivity and design, and was linked to a local database to allow for more efficient data analysis and processing. The only difference from the original was the insertion of a brief interactive questionnaire between each test item for reasons described above. Test takers therefore had a near-identical experience to that of real-world *CAE* CBT test takers.

### Eye tracking: technical specifications

The eye tracker used was a Tobii T60. Unlike most eye trackers these new devices dispense with chin rests, helmets and other distractions, and in addition the tracking cameras are hidden in the monitor casing, ensuring that users' behaviour is as natural as possible without unwarranted intrusion on their mental processing. The T60 sample rate is 60 Hz per second, which allows detailed tracking of normal reading, and it was set to a screen recording rate of 10 frames per second. (Full technical specifications can be found at: www.tobii.com) In addition the device was furnished with binocular tracking (rather than tracking on one eye only), a user camera and speakers for playing the tutorial soundtrack.

### Participants

One hundred and three multinational participants studying at a UK university, representing more than 15 nationalities and language groups and ranging in academic level from pre-university Foundation year students (n=29), to Year 1 (n=41) and Year 2 undergraduate students (n=33), completed the test items from the *CAE* computerised Reading test described below. Ages ranged from 17–20 (n=27, 26.2%), 21–25 (n=71, 68.9%) and 26–35 (n=5, 4.9%).

The test-taking activity of a sample of these (n=35, 36%) was recorded using Tobii screen recording software, which captured every key press, mouse movement, eye movement and facial expression. The sample selected for eye tracking was weighted to ensure good representation across all academic levels, so that the eye-tracking data covered students at Foundation level, Year 1 and Year 2 undergraduate levels. Apart from that, selection was random.

All students signed appropriate ethics forms and personal information forms. In addition they were asked to rate their own familiarity with computers in general and onscreen tests in particular. As was expected with this young and educated group, all reported extensive familiarity with computer technology and onscreen tests of various kinds.

### Test items

The original *CAE* test consisted of six texts and a total of 34 multiple-choice (MC) items. Time constraints and technical constraints (described below) led to the selection of four of these texts, with a total of 13 test items (Parts 1 and 3, with items 1–6 and 13–19).

In the original *CAE* test Part 1 consisted of three short texts with two MC items on each, a total of six items, all of which were included in our test. Part 2 of the original *CAE* test (with questions 7–12) consisted of a task in which test takers drag and drop correct parts of a text into place to complete the whole. This could in principle be eye tracked for each participant, but given the huge variation in scrolling and dragging behaviour it would be complex to compare any two participants' behaviour through an eye-tracking device, so for this reason Part 2 was omitted. Part 3 (items 13–19) consisted of a single long text with a side scrollbar, and although this presented similar analytical problems in terms of comparing eye-tracking behaviour across candidates, it was nonetheless included owing to the importance of testing participants' reading over longer stretches than the short texts in Part 1. With respect to Part 4, given the inadvisability of tracking eye movements over too lengthy a period, it was decided to omit this last section to ensure that the whole test would take no more than approximately 30 minutes.

### Onscreen questionnaire

The retrospective questionnaire, which appeared after each test item was completed, aimed to elicit from participants their own idea of how they had dealt with that item. In terms of content and design it drew on the paper questionnaire used by Weir, Hawkey, Green and Devi's study (2009), but in the light of discussion with two of the authors of that paper (Weir and Green), it was modified and shortened in an attempt to make it clearer.

The version used consisted of three parts. The first asked about how they had approached the text and questions, and offered three options as follows, from which participants had to choose one:

*Before reading the question, I:*

*a – read the text or part of it slowly and carefully.*
*b – read the text or part of it quickly and selectively to get a general idea of what it was about.*
*c – did not read the text.*

Part 2 presented five questions asking about particular cognitive strategies. Participants could choose more than one if they wished:

*To find the answer to the question I tried to:*

*1 – match words that appeared in the question with exactly the same words in the text.*
*2 – match words that appeared in the question with similar or related words in the text.*
*3 – search quickly for part(s) of the text to answer the question.*
*4 – read part(s) of the text slowly and carefully to get the answer to the question.*
*5 – read relevant part(s) of the text again carefully.*

Part 3 presented two options aiming to distinguish between local and global processing. Participants had to choose one.

*I found the answer:*

*6 – within a single sentence.*
*7 – by putting information together across sentences.*

This gave a total of up to seven responses per candidate: a maximum of one in Part 1, five in Part 2 and one in Part 3. The focus of the questionnaire was therefore on various aspects of the processes which the readers had used, aiming to gain insights as to whether they had read globally or locally, carefully or expeditiously, had used word-search strategies for example, had attempted to combine information across sentences and so on.

### Procedure

After all personal information forms, consent forms and computer familiarity forms had been completed, the project proceeded as follows:

*Stage 1*

For those using the eye tracker, participants' individual eye fixations and saccades were carefully calibrated using the Tobii calibration tool, which identifies each person's individual pattern of gaze and saccade behaviour and ensures the accuracy of the subsequent tracking of their reading during the test. This calibration was carried out individually for each participant.

*Stage 2*

Each participant watched a short video tutorial modelled closely on the *CAE* CBT tutorials, explaining each aspect of the process they were about to follow. This video also explained the retrospective questionnaire which appeared between each test item.

*Stage 3*

Participants then completed the *CAE* reading items onscreen. They were given a time indication of 30 minutes for the 13 questions. As noted above, the test experience followed the *CAE* CBT procedures except that immediately after answering each test item participants were presented with an interactive screen eliciting their retrospective recall of the cognitive processes they had used to answer that question. The screen also showed the question itself again, so as to stimulate more accurate recall. All answers and responses were saved to a database.

## Analysis

When the tests had been completed the process of analysis was initiated, which consisted of the following three stages: item selection, participant selection and the analysis of the eye-tracking data.

### Item selection

In order to investigate whether the participants had employed the range of cognitive processing types identified in Khalifa and Weir (2009), as discussed above, the first step in the analysis was to select items from the *CAE* test which covered the range of cognitive skills. To this end the 13 test items were examined by an expert focus group so as to identify the cognitive processing operations which each item aimed to elicit. For example items which were devised so as to test a reader's ability to find and make use of a lexical item, at a lower level of complexity, were distinguished from items devised to test a reader's ability to make connections at a higher, text level, and so on. On this basis five items (5, 13, 17, 18, 19) were selected which covered the range of cognitive processes in Khalifa and Weir's model (2009), from the lowest (at the lexical level, item 18) to the highest (19, drawing on the whole text), as set out in Table 1.

These five items were then analysed on the basis of scores from the whole cohort (n=103) to ensure that they were functioning well, so far as this sample size could tell us, at the appropriate level of difficulty. As can be seen in Table 2, it was confirmed that facility values of the five items fell within the range 0.42–0.63, and discrimination indices of these items were all .25 or greater, both of which Henning (1987) suggests as acceptable ranges of these values respectively.

**Table 1: Specifics of items selected for further analysis**

| **Item number in CAE test** | 5 | 13 | 17 | 18 | 19 |
|---|---|---|---|---|---|
| **Target area of each item** | Across two paragraphs | Within one paragraph | Within one sentence | Particular lexis (within sentence) | Across whole text |
| **Facility value (from n=103)** | 0.52 | 0.63 | 0.42 | 0.62 | 0.60 |
| **Discrimination index (item-total correlation)** | 0.33 | 0.25 | 0.36 | 0.50 | 0.41 |

### Participant selection

The sample of participants whose recordings would be analysed was drawn from those whose *CAE* tests were eye tracked (n=35 out of the original 103). Given the aims of the project, the sample was further restricted to the stronger candidates since our aim was to investigate how the *CAE* items performed when taken by candidates at the appropriate level, and not by candidates below that level. As all the participants had also taken a set of 11 *IELTS* onscreen reading items of different types on the same day this additional yardstick of students' onscreen reading abilities was available, and 'strong' candidates were therefore defined as those who had scored highly on both the *CAE* items and the *IELTS* items combined (i.e. those with more than 50%, of the possible 24). This gave a total pool of 15 participants who were demonstrably proficient onscreen readers in general terms and not only on *CAE* test items in particular, since they had also performed well on the *IELTS* items.

Of these 15 more proficient participants, not all had correctly answered all of the five items selected for analysis, so in addition, for each test item, participants from the pool were identified who had that particular item correct. Apart from item 17, which only four of the pool had answered correctly, six participants were chosen for analysis for each item. (Of course, these were not the same six participants for each item.) The upshot of this was that the eye-tracking data to be analysed consisted of a total of 28 recordings, i.e. the responses of six strong participants who answered correctly for each of four items, and the four participants from the pool who had answered question 17 correctly. In the report on the findings which follows these are given the initials A-F for reasons of anonymity, though again it should be noted that participant A is not the same person in each item analysed.

### Analysis of eye-tracking data

The 28 onscreen recordings were then analysed through both the Tobii Studio software and through detailed visual and statistical analysis. In order to focus this analysis nine questions (set out in Table 2) were posed for each participant and each test item. These questions were designed to examine all key aspects of the readers' processing, including those covered in their online questionnaire, to allow for later comparison. Alongside each question in Table 3 can be seen the approach used or the software tool employed in investigating that question; these will be further explained and exemplified below.

These questions permitted insight into the kinds of cognitive processes which participants had used when successfully answering each test item. For example, item 19 in the *CAE* test requires test takers specifically to read the whole text (the 'target' in our terms), so investigation of the range of questions in Table 2 permitted us to ascertain whether participants had in fact done so. Our approach therefore allowed unprecedented insights into readers' moment-by-moment reading behaviour as they responded

**Table 2: Analysis of eye-tracking data**

| | **Questions** | **Analytical tools** | | | |
|---|---|---|---|---|---|
| | | **Visual analysis of eye movements (video data)** – *see e.g. Figure 1* | **Visual analysis of Gaze Plot data** – *see e.g. Figures 2–4* | **Heat Map data** – *see e.g. Figure 6* | **Automated statistical analysis of fixations** – *see Appendices 1–4* |
| 1 | **Did the participant read the question? (Defined as at least 3 aligned fixations)** | ✔ | ✔ | ✔ | ✔ |
| 2 | **Did the participant read the question BEFORE carefully reading the text?** | ✔ | ✔ | – | – |
| 3 | **Did the participant use expeditious search strategies to locate the correct site of the answer efficiently?** | – | ✔ | – | – |
| 4 | **Did the participant read *all* question options?** | ✔ | ✔ | ✔ | ✔ |
| 5 | **Did the participant read the question options *carefully*? (min. 3 fixations per option)** | ✔ | ✔ | ✔ | ✔ |
| 6 | **Did the participant *skim* options (fewer than 3 fixations)** | – | ✔ | ✔ | ✔ |
| 7 | **(Qs 5, 13, 17, 18) Did the participant focus most heavily on the target area? (see Table 1 for how this was defined for each item)** | ✔ | ✔ | ✔<br>(for non-scrolling items, Q5 and Q13) | ✔<br>(for non-scrolling items, Q5 and Q13) |
| 8 | **Did the participant read *more than one paragraph* carefully?** | ✔ | ✔ | ✔<br>(for non-scrolling items, Q5 and Q13) | ✔<br>(for non-scrolling items, Q5 and Q13) |
| 9 | **(Q 19 only) Did the participant scroll and sample various parts of text?** | ✔ | ✔ | – | – |

to each test item, and unprecedented insight into whether each item was functioning correctly in terms of the cognitive processes it was eliciting – an important part of its validity, as argued above.

Two analysts independently examined the eye-tracking data for each of the 28 onscreen recordings in the light of the nine questions in Table 2, then the analyses were compared. Of the total of 224 judgements made (eight questions x 28) the raters agreed on 213 and disagreed on only 11, an agreement of 95.1%. The high level of agreement is explained by the fact that the eye-tracking data offers a remarkable degree of clarity to the analyst, with few areas of doubt. The 11 disagreements were then resolved through discussion to give the results set out in the Findings section below.

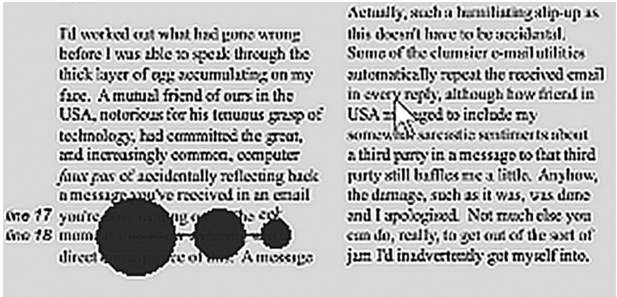
## Analytical tools

Before examining the results it is important for the sake of clarity to explain each of the tools of analysis outlined in Table 2 so as to clarify how they facilitated the analysts' judgements.

### Visual analysis of eye movements (video data)

The Tobii software allows the analyst to follow the moment-by-moment reading of the participant plotted onscreen by a series of lines (indicating saccades) and circles of various sizes (representing fixations – smaller circles for shorter fixations, and larger ones for longer ones. See Figure 1).

**Figure 1: Example of three fixation and two saccade representations**



I'd worked out what had gone wrong before I was able to speak through the thick layer of egg accumulating on my face. A mutual friend of ours in the USA, notorious for his tenuous grasp of technology, had committed the great, and increasingly common, computer *faux pas* of accidentally reflecting back a message you've received in an email you're [illegible] the [illegible] moment [illegible] direct [illegible] of [illegible]. A message

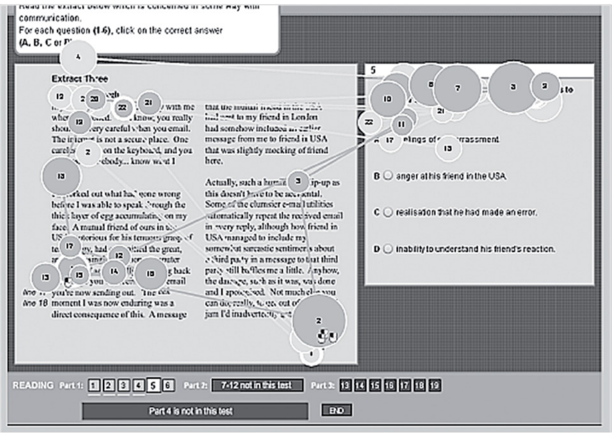
Actually, such a humiliating slip-up as this doesn't have to be accidental. Some of the clumsier e-mail utilities automatically repeat the received email in every reply, although how friend in USA m[illegible]ged to include my somewh[illegible] sarcastic sentiments about a third party in a message to that third party still baffles me a little. Anyhow, the damage, such as it was, was done and I apologized. Not much else you can do, really, to get out of the sort of jam I'd inadvertently got myself into.

This tool allows for the detailed observation and analysis of various aspects of the reading process, since the recording can be slowed for easier observation. However, with this tool it can be difficult to see larger patterns of behaviour, which are better observed with Gaze Plot and Heat Map tools described below.

### Visual analysis of Gaze Plot data

The Gaze Plot tool allows for the analysis of patterns which might be missed on the video, since it illustrates graphically the fixations and saccades of each reader for a selected segment, numbered in order. The Gaze Plot illustrated in Figure 2, for example, shows a comparison between two readers, coloured light and dark respectively, on the same screen, demonstrating the detailed picture which the tool can give of readers' patterns of reading.

**Figure 2: Example of Gaze Plot data, showing two readers' eye movements superimposed**



communication.
For each question (1-6), click on the correct answer
(A, B, C or D).

Extract Three

5

B ○ anger at his friend in the USA

C ○ realisation that he had made an error.

D ○ inability to understand his friend's reaction.

READING Part 1: 1 2 3 4 5 6 Part 2: 7-12 not in this test Part 3: 13 14 15 16 17 18 19

Part 4 is not in this test END

In another example of Gaze Plot data, Figure 3 shows a reader who read two paragraphs of a text, whereas Figure 4 by contrast shows a reader who chose to read only the first paragraph. This is particularly useful when answering question 8 in Table 2 above, to identify how much of each text the participants covered.

**Figure 3: Participant completing question 13 – note the coverage of the whole text**

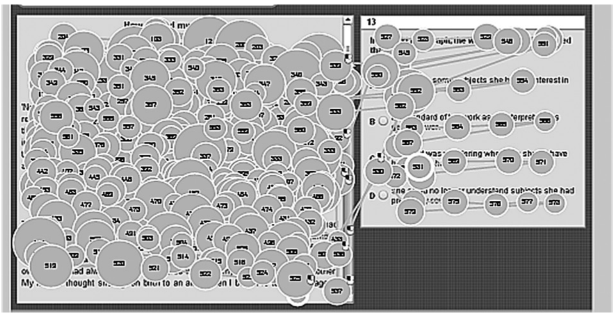
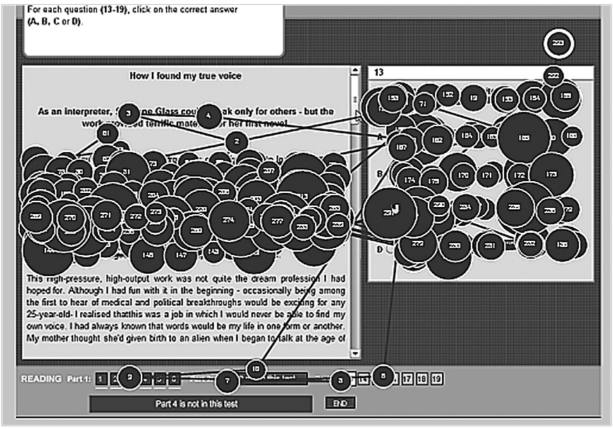


**Figure 4: Participant completing question 13 – note the focus on paragraph one only**



For each question (13-19), click on the correct answer
(A, B, C or D).

How I found my true voice

13

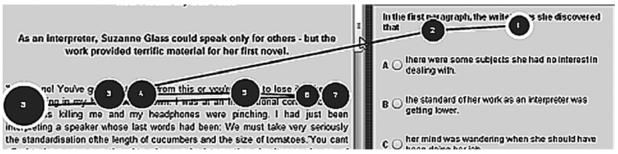
This high-pressure, high-output work was not quite the dream profession I had hoped for. Although I had fun with it in the beginning - occasionally being among the first to hear of medical and political breakthroughs would be exciting for any 25-year-old- I realised thatthis was a job in which I would never be able to find my own voice. I had always known that words would be my life in one form or another. My mother thought she'd given birth to an alien when I began to talk at the age of

READING Part 1:

Part 4 is not in this test END

Figure 5 shows how saccades can identified through Gaze Plot data. This tool is valuable in answering question 3 in Table 2 regarding expeditious reading, since it can show, for example, when the reader uses search reading/expeditious strategies to find the correct part of the text. In Figure 5 the reader has just read the question, then (since the question mentions Paragraph 1) uses expeditious reading skills to find and locate the correct part of the text, jumping from fixation number 2 to the correct part of the text at number 3.

**Figure 5: Example of a saccade indicating expeditious reading after reading the question (the relevant saccade is between fixation numbers 2 to 3)**



As an Interpreter, Suzanne Glass could speak only for others - but the work provided terrific material for her first novel.

no! You've g from this or you'r to lose ing in my onal co s killing me and my headphones were pinching. I had just been interpreting a speaker whose last words had been: We must take very seriously the standardisation ofthe length of cucumbers and the size of tomatoes.'You cant

In the first paragraph, the write s she discovered that

A ○ there were some subjects she had no interest in dealing with.

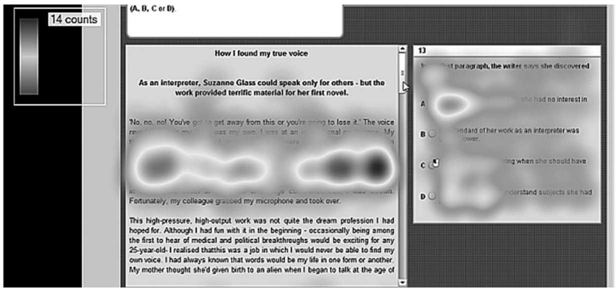
B ○ the standard of her work as an interpreter was getting lower.

C ○ her mind was wandering when she should have

### Visual analysis of Heat Map data

The Tobii Heat Map tool allows for the analysis of frequency and length of fixation in the form of a heat map, giving a view of the areas to which the reader gave the most visual attention. The Heat Map in Figure 6, for example, shows that the reader examined paragraph 1 most closely, and precisely which parts of paragraph 1 they examined, and shows that they also looked at all parts of the question and options. (The original is in colour, which cannot be reproduced here.) In conjunction with the statistical tools (see below) this tool can therefore give a clear sense of the areas to which the participant gave most attention.

**Figure 6: Example of Heat Map data (the original is in colour)**



(A, B, C or D).

How I found my true voice

As an interpreter, Suzanne Glass could speak only for others - but the work provided terrific material for her first novel.

'No, no, no! You've got to get away from this or you're going to lose it.' The voice was my own. I was at an onal My

Fortunately, my colleague grabbed my microphone and took over.

This high-pressure, high-output work was not quite the dream profession I had hoped for. Although I had fun with it in the beginning - occasionally being among the first to hear of medical and political breakthroughs would be exciting for any 25-year-old- I realised thatthis was a job in which I would never be able to find my own voice. I had always known that words would be my life in one form or another. My mother thought she'd given birth to an alien when I began to talk at the age of

13

st paragraph, the writer says she discovered

A she had no interest in

B ○ andard of her work as an interpreter was lower.

C ○ ring when she should have

D ○ nderstand subjects she had

### Automated statistical analysis

The Tobii Studio software facilitates detailed statistical analysis of reader behaviour. Examples are given in Appendices 1–4 to illustrate the kind of data the tool can produce. The illustrations in the appendices are taken from *CAE* question 5; in Appendix 1 is data regarding the number of times each reader fixated on the question itself, while Appendix 2 shows how long it took in seconds before each participant looked at the question. Appendix 3 shows how frequently each reader fixated on each question option in item 5, and Appendix 4 shows how long each reader spent on each option. Here it is noteworthy, for example, that for many of the participants, but not all, option 3 seemed to be more distracting. These illustrations demonstrate the kinds of numerical data which were available in the analysis.

## Findings and discussion

The results of the analysis are summarised in Table 3. Row 1 of the table sets out the questions which were evaluated by the analysts using the tools outlined in Table 2 above. Rows 2–6 set out the results of the analysis for each test item in turn. Row 7 sets out the totals for each question and row 8 sets out the possible maximum rating for each question. Row 9 then sets out the percentages.

The main findings are as follows, for each question:

- It was clear from column A that 100% of participants had read each question carefully (as we would expect of proficient and computer-literate students, though it is worth noting that some less proficient students not examined in this study did not do so).
- From column B it is apparent that all participants on every question bar one read the question before reading the

**Table 3: Summary of analysis**

| Row 1 | Item | Target area | No. of partici-pants | A: Did participants read the question? (at least 3 fixations) | B: Did participants read the question BEFORE reading the text carefully? | C: Did participants use expedi-tious search strategies to locate the correct place of the answer efficiently? | D: Did participants read *all* question options? | E: Did participants read question options *carefully*? (3 fixations per option) | F: Did participants *skim* options (fewer than 3 fixations) | G: (Not Q19) Did partici-pants fixate or focus most heavily on target? | H: Did participants read *more than one paragraph* carefully? | I: (Q19 only) Did partici-pants scroll and/ or sample various parts of text? |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | **Q5** | **Across two para-graphs** | 6 | 6 | 6 | 6 | 6 | 6 | 0 | 3 | 4 | |
| 3 | **Q13** | **Within one paragraph** | 6 | 6 | 5 | 6 | 6 | 6 | 0 | 6 | 2 | |
| 4 | **Q17** | **Within one sentence** | 4 | 4 | 4 | 4 | 4 | 4 | 0 | 4 | 3 | |
| 5 | **Q18** | **Particular lexis (within sentence)** | 6 | 6 | 6 | 6 | 5 | 5 | 1 | 6 | 1 | |
| 6 | **Q19** | **Across whole text** | 6 | 6 | 6 | 4 | 6 | 6 | 0 | | 3 | 3 |
| 7 | | | | 28 | 27 | 26 | 27 | 27 | 1 | 19 | 13 | 3 |
| 8 | **Max** | | | 28 | 28 | 28 | 28 | 28 | 28 | 22 | 28 | 6 |
| 9 | **%** | | | 100.0% | 96.4% | 92.9% | 96.4% | 96.4% | 3.6% | 86.4% | 46.4% | 50.0% |

text (96.4%, the exception being one participant on Q13, column B).

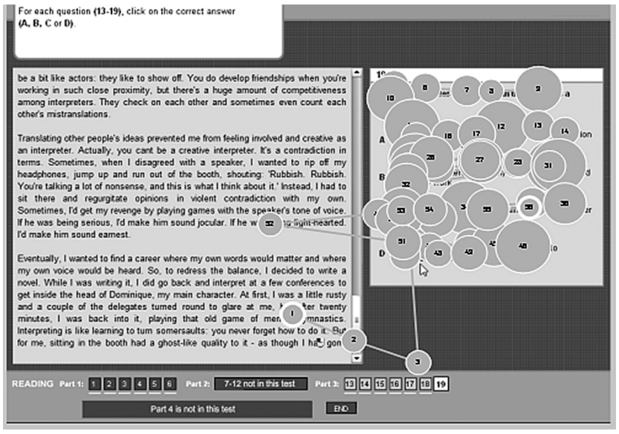
- Column C shows that a sizeable majority (92.9%) used appropriate expeditious strategies to find the correct part of the text for each answer.
- In column D we see that every participant read all the options on all test items except in one instance (i.e. Q18 in column D).
- Column E shows that all participants (96.4%) had read all the options carefully, with one exception where the participant had merely skimmed through one option. This can also be seen in column F.
- Column G is the most pertinent to this study, since it shows that in almost all cases participants had focused on the area targeted by the test item, meaning the items have a strong claim to cognitive validity. It is worth examining this column in some detail:
- Column G, row 2 shows that half of the six students answering question 5 focused entirely on the two target paragraphs to get the answer, as anticipated. However, one student read the whole text carefully instead, and the remaining two merely skimmed through the first paragraph (with only three and 13 fixations respectively, lasting only 0.75 and 3.01 seconds) and then focused heavily on the second paragraph. Given that they all answered this item correctly this suggests that these two candidates obtained enough information in one paragraph to satisfy them, or else were fortunate.
- Column G, row 3 shows that all students performed on question 13 as anticipated, focusing heavily on the target paragraph as a whole. Figure 6 is taken from this question, and shows vividly in graphic form precisely where the student focused attention.
- Column G, row 4 shows that all students performed on question 17 as anticipated, focusing heavily on the target sentence.
- Column G, row 5 shows that all students performed on question 18 as anticipated, focusing heavily and repeatedly on the target lexis. The mean fixation duration on the relevant lexis was 1.82 seconds for all participants, with an average fixation count of 3.17. If we compare this with another randomly selected piece of lexis from the same paragraph, which received a mean of 0.13 seconds of attention and a mean of 0.66 of fixation counts, it is clear that the target lexis was successfully identified and received a high level of focus among these proficient test takers. This in turn implies that the item was successfully targeting the appropriate cognitive processing activity (focusing on lower level lexical areas).
- Column H also merits careful consideration for each test item:
- Column H, row 2 shows again, as discussed above, that for question 5 two students did not read the two paragraphs fully, for reasons already discussed.
- Column H, row 3 shows that although the target answer for question 13 was to be found within one paragraph, two students nevertheless read more than that one paragraph. The other four participants were highly focused in their reading – an example can be seen in Figure 6. The two who read more than necessary were presumably checking that their expeditious search reading had worked properly, and that they had not missed anything.
- Column H, row 4 implies also that most students (three out of four) also read beyond the target section, reading more than one paragraph carefully even though the answer was found within a particular sentence. Since they had already all used expeditious skills to find the correct sentence, as seen in column C, this was presumably for checking.
- Column H, row 5 suggests that as all students quickly found the correct lexis to answer the question, they did not need (except one) to read any other paragraphs, again a sign of their efficiency and confidence as readers.
- Column H, row 6 concerns question 19 which required a grasp of the whole text. It was interesting to see different strategies for this question. Three students, as is clear from Column H, row 6, read more than one paragraph carefully, but three did not – in fact they read almost nothing before identifying the correct answer, presumably because by that stage they had already built up a sufficient idea of the whole text to choose the correct response.
- To illustrate this further, Table 4 shows the amount of time spent by each of the six participants on question 19, on the text and the question/options respectively. This shows – perhaps surprisingly – that all participants apart from B spent longer on the questions than on the text, a mean of 30.57 seconds on the former and 16.57 seconds on the

**Table 4: Comparison of time (seconds) spent on Q19 *text* versus Q19 *questions***

| Participant | Total Visit Duration Q19 Questions (Mean) (seconds) | Total Visit Duration Q19 Questions (Sum) (seconds) | Total Visit Duration Q19 Text (Mean) | Total Visit Duration Q19 Text (Sum) |
|---|---|---|---|---|
| A | 29.43 | 29.43 | 19.51 | 19.51 |
| B | 41.39 | 41.39 | 47.06 | 47.06 |
| C | 15.7 | 15.7 | 15.14 | 15.14 |
| D | 41.89 | 41.89 | 16.67 | 16.67 |
| E | 35.29 | 35.29 | 0.71 | 0.71 |
| F | 19.73 | 19.73 | 0.32 | 0.32 |
| **All recordings** | 30.57 | 183.42 | 16.57 | 99.41 |

latter. Some participants spent almost no time at all on the text (e.g. E took 0.71 seconds, and F took 0.3 seconds) which strongly suggests that they had already constructed a strong and confident sense of the text's overall sense. This is graphically illustrated in Figure 7, which shows participant E's eye movements, concentrating heavily on the questions and almost not at all on the text before answering.

**Figure 7: Participant E's eye movements on question 19**



For each question **(13-19)**, click on the correct answer **(A, B, C or D)**.

be a bit like actors: they like to show off. You do develop friendships when you're working in such close proximity, but there's a huge amount of competitiveness among interpreters. They check on each other and sometimes even count each other's mistranslations.

Translating other people's ideas prevented me from feeling involved and creative as an interpreter. Actually, you cant be a creative interpreter. It's a contradiction in terms. Sometimes, when I disagreed with a speaker, I wanted to rip off my headphones, jump up and run out of the booth, shouting: 'Rubbish. Rubbish. You're talking a lot of nonsense, and this is what I think about it.' Instead, I had to sit there and regurgitate opinions in violent contradiction with my own. Sometimes, I'd get my revenge by playing games with the speaker's tone of voice. If he was being serious, I'd make him sound jocular. If he was light-hearted. I'd make him sound earnest.

Eventually, I wanted to find a career where my own words would matter and where my own voice would be heard. So, to redress the balance, I decided to write a novel. While I was writing it, I did go back and interpret at a few conferences to get inside the head of Dominique, my main character. At first, I was a little rusty and a couple of the delegates turned round to glare at me, but after twenty minutes, I was back into it, playing that old game of mental gymnastics. Interpreting is like learning to turn somersaults: you never forget how to do it. But for me, sitting in the booth had a ghost-like quality to it - as though I had gon

READING Part 1: 1 2 3 4 5 6 Part 2: 7-12 not in this test Part 3: 13 14 15 16 17 18 19

Part 4 is not in this test END

Returning to the full summary in Table 3, column 1 relates only to question 19, and shows again the fact that three participants used the scrollbar and read back through the text, reading carefully through several paragraphs, while (as noted above) three others scarcely read the text at all. This is an interesting finding, since rather than reflecting badly on the test item it demonstrates that with items testing global understanding some candidates might adopt a more careful approach, selecting to re-read some parts expeditiously and read certain passages carefully, whereas other participants might already be clear and confident enough not to need to re-read any of the text at all. Both behaviours can be characteristic of proficient readers, and both imply higher level cognitive processing skills.

## Student questionnaires evaluated

As noted in the Methodology section, participants were asked after completing each item to report retrospectively on their recently completed processing operations. When the eye-tracking data had been analysed in detail, as discussed above, it was then possible to compare the eye-tracking data with this participant questionnaire data, and then to compare the two.

Data from the student questionnaires was therefore examined alongside the data gathered from eye tracking, discussed above. In total there were seven questionnaire options for each test item, and a total of 28 eye tracking recordings, giving 196 responses. The participants' responses were then examined in the light of the eye-tracking data and marked as accurate or inaccurate. For example if a participant said she had read the text before reading the question this could easily be checked against the eye-tracking data. If the student did not in fact do so then her response would be marked as inaccurate. To take another example, if a participant responded by saying she had not read the text carefully but the eye-track data suggested otherwise, then that answer too was adjudged inaccurate. The analysis was carried out by two adjudicators independently with an agreement ratio of 88%. Doubtful cases were discussed and agreement reached.

It was found that of the 196 possible choices, participants had been accurate in their self-report in 134 (68.4%) cases and inaccurate in 62 (31.6%) cases. This could be cause for celebration, in that a clear majority of the participants' self-assessments were accurate, but given that their retrospective feedback was elicited immediately after having completed each test item they could surely be expected to be more aware of what they had just been doing. It could therefore be argued that the fact that their accuracy in retrospection is so low casts doubt upon studies which depend heavily on retrospective reporting for gaining insights into cognitive processing.

There are other partial explanations for these results. It is possible that the wording of some parts of the questionnaire confused some participants, which may explain why many stated that they had read the text before the question when they had clearly done the opposite. It is also possible that participant fatigue played a part. Nonetheless, since these aspects cannot in themselves account for such a high number of inaccurate self-reports it would appear that retrospective reports in cognitive processing research could be less reliable than has often been supposed, and that eye tracking could offer a more reliable guide to cognitive processing in future research into reading.

## Conclusion

This project has researched a set of *CAE* onscreen test items with a view to investigating their cognitive validity. Through the onscreen testing of 103 students, the eye tracking of 36% of them as they completed the test, and then the selection of a sample of more proficient onscreen readers for more detailed analysis, we have shown that the items analysed performed effectively in terms of eliciting from test takers both the *range* of cognitive processing identified in Khalifa and Weir's model (2009) and also the different *levels* of processing from lower areas to more complex levels, including whole text comprehension.

Detailed analysis of each item through a variety of approaches, using the graphic, video and statistical tools afforded by eye-tracking software, as well as careful visual analysis, demonstrated the ways in which these test items were performing in terms of the cognitive processes they were requiring of readers. The set of items together demonstrably tested cognitive processing at the lower levels (e.g. of lexis), the sentence level, the paragraph level, across paragraphs and at whole-text level. The set of items can therefore claim with some confidence to have cognitive validity in Khalifa and Weir's terms.

In addition, the research demonstrated the value of using eye tracking to assist in the validation of test items and the possible limitations of traditional retrospective reports on participants' cognitive processes. In our view this research opens exciting new windows, to continue the metaphor, onto both the cognitive processes of readers under test conditions and also onto the ways in which test items can perform when eliciting particular cognitive processes in reading.

## References and further reading

Afflerbach, P and Johnston, P (1984) On the use of verbal reports in reading research, *Journal of Reading Behaviour* 16 (4), 307–321.

Alderson, J C (2000) *Assessing Reading*, Cambridge: Cambridge University Press.

Anson, C, Rashid Horn, S and Schwegler, R (2009) The Promise of Eye-Tracking Methodology for Research on Writing and Reading, *Open Words: Access and English Studies* 3 (1), 5–28.

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

Bertram, R (2011) Eye movements and morphological processing in reading, *The Mental Lexicon* 6 (1), 83–109.

Buscher, G, Biedert, R, Heinesch, D and Dengel, A (2010) Eye-tracking analysis of preferred reading regions on the screen, in Mynatt, E D, Schoner, D, Fitzpatrick, G, Hudson, S E, Edwards, W K and Rodden, T (Eds) *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Extended Abstracts Volume*, Atlanta, Georgia, USA, 10–15 April, 2010, 3,307–3,312.

Carver, R (1992) Reading Rate: theory, research and practical implications, *Journal of Reading* 36 (2), 84–95.

Cordon, L and Day, J (1996) Strategy use on standardized reading comprehension tests, *Journal of Educational Psychology* 88, 288–95.

de Greef, T, Botzer, A and Van Maanen, P-P (2010) Eye-tracking = Reading the mind, in *Proceedings of the 28th Annual European Conference, on Cognitive Ergonomics (ECCE 2010)*, Delft, Netherlands, August 25–27 2010, New York, NY: ACM Press, 303–304.

Duchowski, A (2002) A Breadth-First Survey of Eye-tracking Applications, *Behavior Research Methods, Instruments, and Computers (BRMIC)* 34 (4), 455–470.

Eger, N, Ball, L, Stevens, R and Dodd, J (2007) Cueing Retrospective Verbal Reports in Usability Testing Through Eye-Movement Replay, in *Proceedings of HCI 2007, The 21st British HCI Group Annual Conference, University of Lancaster, UK.* Available online http://www.bcs.org/server.php?show=ConWebDoc.13300

Gass, S and Mackey, A (2000) *Stimulated Recall Methodology in Second Language Research*, Mahwah, NJ: Lawrence Erlbaum.

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

Henning, G (1987) *A Guide to Language Testing: Development, Evaluation, Research*, Cambridge, MA: Newbury House.

Hoover, W A and Tunmer, W E (1993) The components of reading, in Thompson, G B, Tunmer, W E and Nicholson, T (Eds) *Reading acquisition processes*, Clevedon, UK: Multilingual Matters Ltd, 1–19.

Khalifa, H and Weir, C (2009) *Examining Reading: Research and practice in assessing second language reading*, Studies in Language Testing, volume 29, Cambridge: UCLES/Cambridge University Press.

Pressley, M, and Afflerbach, P (1995) *Verbal protocols of reading: the nature of constructively responsive reading*, Hillsdale, NJ: Lawrence Erlbaum.

Rayner, K (1998) Eye movements in reading and information processing: 20 years of research, *Psychological Bulletin* 124 (3), 372–422.

Rayner, K and Pollatsek, A (1989) *The psychology of reading*, Englewood Cliffs, NJ: Prentice Hall.

Ruiz-Primo, M, Shavelson, R, Li, M and Schultz, S (2001) On the Validity of Cognitive Interpretations of Scores From Alternative Concept-Mapping Techniques, *Educational Assessment* 7 (2), 99–141.

Salvucci, D and Goldberg, J (2000) Identifying fixations and saccades in eye-tracking protocols, in *Proceedings of the Eye-tracking Research and Applications Symposium*, Palm Beach Gardens, FL, USA, 6–8 November 2000, New York: ACM Press, 71–78.

Spivey, M, Richardson, D and Dale, R (2009) The movement of eye and hand as a window into language and cognition, in Morsella, E and Bargh, J (Eds) *Oxford Handbook of Human Action*, New York: Oxford University Press, 225–248.

Urquhart, A and Weir, C (1998) *Reading in a second language: process, product and practice*, London: Longman.

Vipond, D (1980) Micro- and macro-processes in text comprehension, *Journal of Verbal Learning and Verbal Behaviour* 19, 276–296.

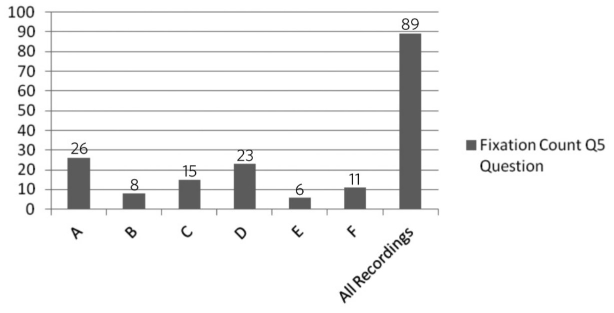
Weir, C (2005) *Language Testing and Validation: an evidence based approach*, Basingstoke/New York: Palgrave Macmillan.

Weir, C, Hawkey, R, Green, T and Devi, S (2009) The cognitive processes underlying the academic reading construct as measured by IELTS, in *British Council/IDP Australia IELTS Research Reports* volume 9 (4), 157–189. Available online from http://www.ielts.org/researchers/research.aspx

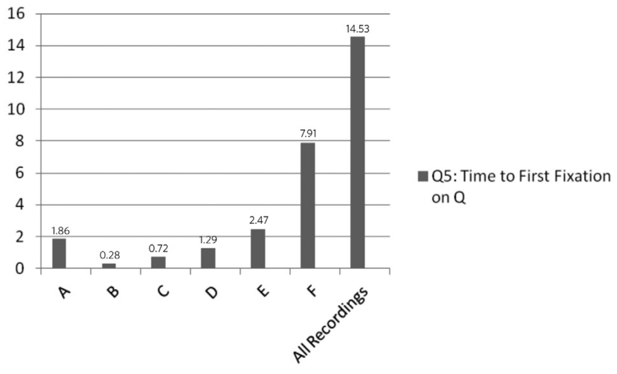
## Appendices

Appendices 1–4 illustrate the kind of data the tool can produce. The illustrations in the appendices are taken from *CAE* question 5; Appendix 1 shows the number of times each reader fixated on the question itself, while Appendix 2 shows how long it took in seconds before each participant looked at the question. Appendix 3 shows how frequently each reader fixated on each question option in item 5, and Appendix 4 shows how long each reader spent on each option. Here it is noteworthy, for example, that for many of the participants, but not all, option 3 seemed to be more distracting. These illustrations demonstrate the kind of numerical data which was available in the analysis.

**Appendix 1: Number of fixations on question 5**

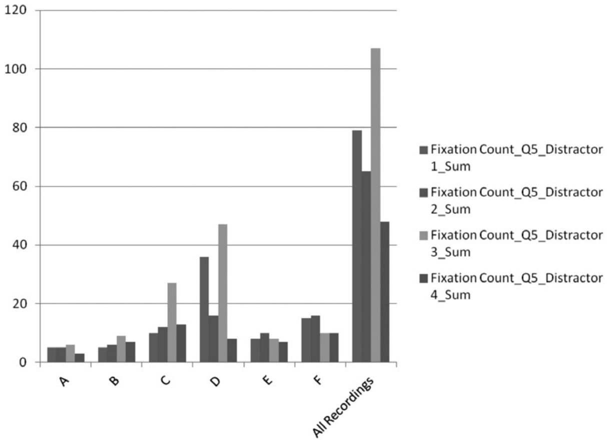


**Appendix 2: Time to first fixation on question 5**



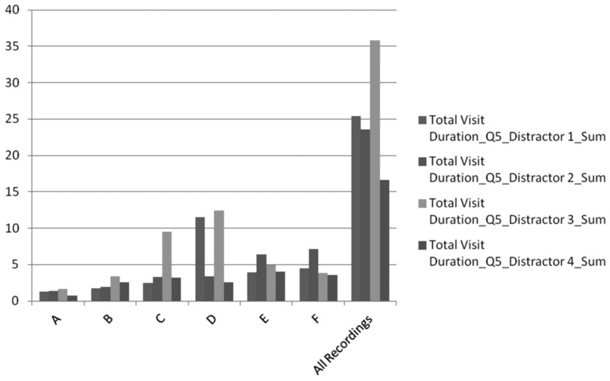
**Appendix 3: Number of fixations on question 5 options**

| Participant | Fixation count | | | |
|---|---|---|---|---|
| | Option 1 (correct response) | Option 2 | Option 3 | Option 4 |
| A | 5 | 5 | 6 | 3 |
| B | 5 | 6 | 9 | 7 |
| C | 10 | 12 | 27 | 13 |
| D | 36 | 16 | 47 | 8 |
| E | 8 | 10 | 8 | 7 |
| F | 15 | 16 | 10 | 10 |
| **All Recordings** | **79** | **65** | **107** | **48** |



**Appendix 4: Visit duration for question 5 MC options (amount of time in seconds spent on each option by each participant)**

| Participant | Total visits (seconds) | | | |
|---|---|---|---|---|
| | Option 1 (correct response) | Option 2 | Option 3 | Option 4 |
| A | 1.32 | 1.4 | 1.68 | 0.75 |
| B | 1.78 | 1.93 | 3.36 | 2.53 |
| C | 2.43 | 3.33 | 9.53 | 3.16 |
| D | 11.51 | 3.38 | 12.42 | 2.56 |
| E | 3.9 | 6.38 | 4.97 | 4.06 |
| F | 4.46 | 7.13 | 3.85 | 3.55 |
| **All Recordings** | **25.4** | **23.55** | **35.82** | **16.62** |



# Investigating figurative proficiency at different levels of second language writing

**JEANNETTE LITTLEMORE** UNIVERSITY OF BIRMINGHAM, UK
**TINA KRENNMAYR** VU UNIVERSITY AMSTERDAM, THE NETHERLANDS
**JAMES TURNER** UNIVERSITY OF BIRMINGHAM, UK
**SARAH TURNER** UNIVERSITY OF BIRMINGHAM, UK

## Introduction

In very broad terms, metaphor involves describing one thing in terms of another (e.g. when women's careers are described as 'hitting a glass ceiling'). Metonymy involves the use of one entity to refer to a related entity (e.g. the use of the term 'Hollywood' to refer to the US film industry). Studies of metaphor (and to a lesser extent metonymy) have shown that they perform key functions, such as: the signalling of evaluation; agenda management; mitigation and humour; technical language; reference to shared knowledge; and topic change (Semino 2008). An ability to use them appropriately can thus contribute to a language learner's communicative competence (Littlemore and Low 2006 a and b), and is therefore likely to be a key indicator of a language learner's ability to operate at different levels of proficiency as defined by the Common European Framework of Reference for Languages (CEFR). The CEFR, which forms part of a wider European Union initiative, is a series of descriptions of language abilities which can be applied to any language and can be used to set clear targets for achievements within language learning. It has now become accepted as a way of benchmarking language ability all over the world. There are six levels (A1, A2, B1, B2, C1 and C2). Each level contains a series of Can Do statements, which describe the various functions that one would expect a language learner to perform in reading, writing, listening and

This article has been accepted for publication in Applied Linguistics, published by Oxford Univesity Press.

speaking, at that level. The Can Do statements for writing ability (the focus of this study) can be found at: www.coe.int/t/DG4/Portfolio/?L=EandM=/documents_intro/Data_bank_descriptors.html

In these statements there is a clear progression in terms of the complexity of functions that a learner is expected to perform and we might thus expect their use of metaphor to both change and increase across the different levels. For example, at Level A1, learners are expected to be able to 'write a short, simple postcard, for example sending holiday greetings and fill in forms with personal details'. We would expect very little use of metaphor here, except perhaps for the odd metaphorically used preposition, whereas at Level C2, learners are expected to be able to 'write clear, smoothly-flowing text in an appropriate style. . . write complex letters, reports or articles which present a case with an effective logical structure which helps the recipient to notice and remember significant points [and] . . . write summaries and reviews of professional or literary works'. Here we would expect learners to use metaphor to convince and persuade as well as to link their ideas to one another. To date, there has been no detailed investigation into how a learner's use of metaphor develops across these different levels. Nor has there been any investigation into the ways in which a learner's L1 background influences their use of metaphor and metonymy at different levels of proficiency in learners' writing. Such a study would be useful as it could contribute descriptors pertaining to the use of metaphor and metonymy which could then be used in training materials. The findings of such a study would also be useful for organisations, such as Cambridge ESOL, which are involved in language assessment, as they could be incorporated into the marking criteria for their written examinations.

In this article we describe a study, funded by Cambridge ESOL, which used the Cambridge Learner Corpus (CLC), a unique corpus of exam scripts at each of these levels, to meet the following aims:

- to identify features of metaphor that distinguish the different CEFR levels, as measured by the Cambridge exams
- to provide descriptors relating to metaphor use that could be incorporated into the different CEFR descriptors for each level of writing for English.

The focus of the study was on metaphor as this has reasonably robust identification technique. It also looked to some extent at metonymy but for reasons mentioned below, our quantitative findings from the metonymy part of the study are not reported here. We limited our study to the top five levels (A2-C2) after observing that virtually no metaphor was produced at Level A1. Our first objective was to measure the amount of metaphor produced across CEFR levels A2 to C2. The most widely used maximally inclusive approach to metaphor identification is the Pragglejaz Group (2007) metaphor identification procedure (MIP). This procedure involves identifying as metaphor *any* lexical unit that has the potential to be processed metaphorically. The analyst begins by identifying all the lexical units in the text (in most, but not all cases, a 'lexical unit' refers to a 'word', but see the Methodology section below). Then for each lexical unit, they establish its meaning in context and decide whether it has a more basic contemporary meaning in other contexts and if so, whether its meaning in the text can be understood in comparison with this more basic meaning. In the majority of cases, the decision was taken to regard a single word as comprising the lexical unit, even when the analyst's intuition might be to class certain uses as phrases, or a dictionary might record two or more words as making up a phraseological unit. The reasoning behind this decision is outlined in Section 3. Basic meanings tend to be more concrete, related to bodily action, or more precise. If this is the case then the lexical unit is marked as being 'metaphorically used'. We used a slightly adapted version of this technique inspired by Steen, Dorst, Herrmann, Kaal, Krennmayr and Parma's MIPVU (2010). Some useful features of the MIPVU for our particular project are that it includes 'direct metaphors' (i.e. similes and the like) as well as 'implicit metaphors', such as the use of 'this' and 'that' or pronouns such as 'it' or 'one' to refer back to metaphorically used words (e.g. The *path* she took was indeed the right *one*), and 'possible personifications' (such as 'the department needs to *act*'). All of these features have been found to vary across languages, and present considerable challenges to learners. However, it treats phrasal verbs and multiword items as single units for analysis. Language learners often make mistakes *within* phrasal verbs and multiword items, suggesting that they may not always be learning them as fixed phrases, and that they may at times be treating them as novel compounds. In order to get at these items we therefore elected to split any phrasal verbs and multiword items whose meanings were deemed to be partially motivated by the basic senses of their constituents. We also included items that involved a change in word class, so 'snaked' would count as a metaphor, even though it has a different word class in its basic sense. Here we follow Deignan's (2005) work, which shows that metaphorical senses often differ formally from their literal counterparts. The technique throws up items that some people might not consider to be metaphor. For example, in our data, the word 'in' in the following sentence:

1 men in the really high positions[1]

would be marked as metaphor because it contrasts and can be understood in comparison to its more basic spatial meaning (inside, a container, room, building etc.). For some analysts, marking this use of 'in' as metaphor would be somewhat counter-intuitive, as it is the most conventional way of expressing this concept and it is very difficult to think of an alternative. It is clearly very different from the use of the term 'black hole' in the following sentence, which also comes from our data:

2 managers tend to fall in a black hole when they retire

The MIP does not make any claims about whether the lexical unit is actually *processed* as a metaphor, only identifies lexical

[1] Metaphorically used lexical units are indicated by solid underlining whereas metonymically used lexical units are indicated by italics. In our examples, only those metaphors and metonymies that are relevant to the particular point that we are making are underlined.

units that have the potential to be processed as metaphor. This is important for studies of metaphor used by language learners, as prepositions may be used in different ways in the learner's own language, a fact which makes their metaphoricity in the target language much more apparent (Littlemore and Low 2006b). For example, the corresponding sentence in Russian would be something like:

3 *мужчины* ***на*** *высоких позициях/постах* --> men **on** high positions[2]

so the metaphoricity of the 'in' may in fact be more salient for a Russian learner of English than it is for a native speaker, who may be less sensitive to the metaphoricity underlying conventional expressions. We also began to develop a methodology for metonymy identification, based on a system proposed by Biernacka (forthcoming), but because this technique is still under development, we do not report our findings in this article.

Most 'dead' or perhaps more appropriately termed 'sleeping' metaphors (Müller 2008) tend to be found within the category of closed-class items, and most 'novel' or 'creative' metaphors tend to involve open-class items. It is therefore interesting to look at how learners make use of open and closed-class items respectively as they may reflect different ways of using metaphor. Our second objective was therefore to explore the extent to which the use that learners make of open-class metaphorical items resembles that which they make of closed-class metaphorical items across the different CEFR levels.

It has been observed that native speakers of English tend to produce metaphor in clusters, that these clusters serve important communicative functions (Cameron and Low 2004), and that some of the most communicatively effective clusters are those that contain mixed metaphors, despite the fact that traditional writing guides often tell writers to avoid mixing their metaphors (Kimmel 2010). One would also expect some development in the production of metaphor clusters in learner writing at the different levels. The third objective was therefore to look at the size, the distribution and the nature of the metaphor clusters produced by learners at each of the levels.

It is important to look not just at the amount of metaphor that is being used but at what learners use metaphor for in their writing, in other words, what functions it is being used to perform. Our fourth objective was therefore to assess the ways in which the learners' use of metaphor contributes to a learner's ability to perform the relevant functions at each of the CEFR levels. We looked at metaphors that occurred both within and outside clusters.

As well as discovering how much metaphor the learners use at each level and what they use it for, it is worth investigating (for teaching purposes) the extent to which they are able to use it accurately. If learners are particularly likely to use metaphor inaccurately at one of the levels, this is useful for teachers to know as they can then address the issue at that particular level. It might also be the case that when learners try out new metaphorical expressions, they use them inaccurately at first, and then develop accuracy at a later stage. It is therefore useful to know if there is a particular stage of learning at which they start to do this as teachers and examiners could then be more lenient in their error marking to allow for experimentation. One might also expect metaphor errors to be due, to some extent, to L1 influence, and one might expect the amount of L1 influence to decrease gradually across the different levels as the learners acquire an understanding of the ways in which metaphor is used in the target language. Alternatively, as Kellerman (1987 a and b) has shown for idioms, L1 influence in metaphor use may peak at the beginning and advanced stages of learning. The fifth objective of our study was to explore the extent to which the use of metaphor in the transcripts appeared to be influenced by the L1 background of the learners, at each of the levels.

## Research questions

The objectives listed above translate into the following research questions:

In two sets of Cambridge ESOL exam scripts (one produced by Greek-speaking learners of English and one produced by German-speaking learners of English):

1. In what ways does the amount of metaphor produced vary across CEFR Levels A2 to C2?
2. In what ways does the use that learners make of open-class metaphorical items resemble or differ from that which they make of closed-class metaphorical items across the different CEFR levels?
3. In what ways does the distribution of metaphor clusters vary across CEFR Levels A2 to C2?
4. In what ways do the functions performed by the metaphor clusters vary across CEFR Levels A2 to C2 and how closely do these functions relate to the CEFR descriptors?
5. To what extent do learners use metaphor 'incorrectly' and how is their use of metaphor influenced by their L1 background?

## Methodology

One hundred essays written by Greek learners of English (20 at each level) and 100 essays written by German learners of English (20 at each level) were selected from the Cambridge English exams (*KET*, *PET*, *FCE*, *CAE* and *CPE*) in the Cambridge Learner Corpus. As far as possible, attempts were made to extract essays on related subjects in order to minimise the impact of topic type in our results[3]. We therefore used the same search terms to extract essays from the corpus at each of the five levels. We chose the words 'politician', 'politics', 'government', 'economy', 'measures' and 'environment'. Search terms such as these reflect domains that have been shown to involve a substantial amount of metaphor (Semino 2008). They are also broad enough to

[2] We would like to thank Anna Eyngorn for this translation.

[3] See www.CambridgeESOL.org/exams for details of the range of topics and format of each exam studied here.

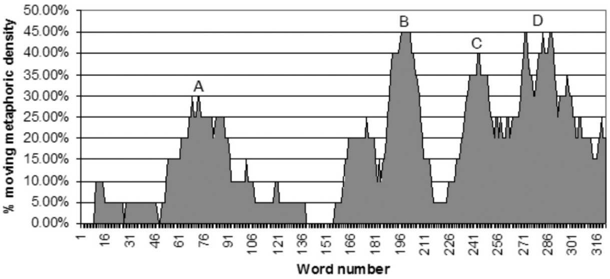
encompass a wide variety of essays, allowing us to extract sufficient data at each of the levels. Because the different CEFR descriptors involve the ability to perform very different functions, the genres of the essays that students are asked to write for the Cambridge ESOL examinations vary considerably across the different levels. At the different levels, students are asked to produce a range of different genres, including letters, emails, narratives, as well as argumentative essays. Thus the term 'essay' is interpreted very broadly in this research project. In our data, the A2 essays consisted entirely of letters that were written to serve very basic transactional or descriptive functions, such as making arrangements to attend an imaginary class, making holiday arrangements, describing objects or recent purchases. The B1 essays also included a large number of letters but they required more evaluation and thus included topics such as descriptions of exciting events at school, giving advice on dilemmas and describing a birthday party. There were also a small number of short stories in our data set at this level. The essays at B2 level were more likely to take the form of argumentative essays or other types of evaluative and/or persuasive writing, such as newspaper articles. They included polemical topics such as the environment, media intrusion, inventions, the importance of foreign languages and the benefits/drawbacks of public transport and the car. At C1 level, there was a wider variety of genres, designed to elicit persuasive and evaluative language, and the essays included nominations for awards, descriptive, discursive, persuasive and comparative academic articles. At C2 level, the essay prompts required the writers to produce and marshal complex arguments in favour of particular actions or to show a deep understanding of abstract concepts. The genres were even more mixed, including award nominations for people and organisations, proposals for urban development, letters of complaint, discursive, comparative and persuasive academic articles, and philosophical treatises on the value of education. These different genres are a good reflection of the range of functions that learners are supposed to be able to perform at each CEFR level.

The essays were then divided into lexical units and entered into an Excel spreadsheet, with one lexical unit on each line. As we saw above, the decision was taken to regard a single word as comprising the lexical unit, even when the analyst's intuition might be to class certain uses as phrases, or a dictionary might record two or more words as making up a phraseological unit (for example, 'grow up'). Studies of second language written and spoken production have shown that language learners often use the wrong verb/preposition/particle combinations in units such as these (Alejo 2010). These findings indicate that learners may at times treat such chunks in a more compositional way than native speakers (NSs), relying on what Sinclair (1991) refers to as the 'open choice' principle as they lack sufficient collocational knowledge to employ the 'idiom principle'. Thus, although certain combinations may have the status of phrases for lexicographers, linguists, or NSs generally, we cannot make any assumptions about their status as phrases for NNSs (see MacArthur and Littlemore, forthcoming for an in-depth discussion of this issue).

In order to identify all potentially metaphorically used lexical units in the essays, we used an adapted version of the MIPVU Metaphor Identification Procedure (Steen et al 2010), which is based on the Pragglejaz Group's (2007) Metaphor Identification procedure introduced above. We also attempted to use a similar technique to identify metonymy in which we looked for contiguity rather than comparison between the basic sense of the lexical unit and its meaning in context (Biernacka, forthcoming). However, because this technique is still under development, the results from this strand of the research are not reported. The metaphors were then categorised into open and closed-class items. We used these figures to calculate the proportions of metaphor used at each level and the proportions of metaphors that comprised open and closed-class items at each level.

A search for metaphor clusters was then conducted using a time series analysis. This technique is normally used to chart the movements of stock prices over time, appearing in the financial section of a newspaper, and the same principle can be used to calculate the 'moving metaphoric density' of a span of discourse. To calculate it, a span size of, say, 20 words is selected. The metaphoric density across the words in this span (words 1 to 20) is calculated. This is equal to the number of items identified as metaphor divided by 20 (the number of items). The result is placed at the mid-point (the 10th word). The span is shifted one word down, and the metaphoric density calculated for the next 20-word span (2 to 21). The result is placed at the mid-point (the 11th word). The metaphoric density of the next span (words 3 to 22) is calculated and placed at the mid-point (the 12th word), and so on until the end of the text is reached. The technique allows the researcher to produce metaphoric density charts, such as the following:

**Figure 1: Illustration of a moving metaphoric density chart for a CAE essay written by a German learner of English**



The metaphor cluster that appears at point B in the above chart was as follows:

4 If a girl develops in a way to like dolls and languages and hate computer and maths this is just fine – but one should not 'push' her in any direction. This widely spread pattern of thinking is mirrored in German politics (German learner of English, *CAE*: C1)

This is a useful graphical technique for identifying metaphor clusters within discourse. It can inform qualitative analysis, by allowing the researcher to identify stretches of text with high localised metaphoric density (i.e. clusters). The next stage was to decide what percentage of metaphor to use as a 'cut-off' point in our definition of a metaphor cluster. Previous studies (e.g. Cameron and Stelma 2004) have used the 'sudden onset' of metaphor as their main identification criterion for a metaphor cluster. Under this approach, the spike that appears at 121 words in Figure 1 would be a candidate for consideration as a metaphor cluster because it follows a long period of relatively low-level metaphor use, even though the actual metaphoric density of this spike is relatively low (10%). However in our study, we wanted to compare the use of metaphor clusters across levels, so we needed to identify a standard starting point in terms of metaphoric density. In order to do this, we conducted manual examinations of the metaphoric density charts for a number of essays at each of the five levels, and analysed them alongside the essays themselves. We looked at clusters at 5% intervals until we reached a level where (a) we could discern visible metaphor use above and beyond the sorts of highly conventionalised metaphorical uses of prepositions and the like, and (b) the number of clusters was not so great as to be meaningless. We agreed that the most 'meaningful' level to start at was 30%, so we looked at clusters of 30%, 35%, 40%, 45%, 50% and 55+% (there were very few clusters at this level so it made no sense to look for clusters of 60% and above). Taking 30% density as our stating point, we then calculated the number and distribution of clusters that appeared at each level in both data sets.

The number of clusters produced at each level was calculated, using this technique, and measures were made of the densities of the clusters. We then conducted a manual search of the metaphors that appeared both within and outside the clusters to establish how learners were using metaphor at each of the levels. We focused both on what they were doing with the metaphor and on what functions it was being used to perform at each level.

In order to establish the percentage of errors that involved metaphor and to assess the role of L1 influence in these errors, we took 25 essays (five from each level) from the German-speakers' corpus and coded them for error according to two marking criteria: a 'strict' criterion under which non-native-like phraseology (e.g. 'all the world' instead of 'the whole world') was counted as wrong and a 'generous' criterion, under which non-native-like phraseology was counted as correct. We then had a native speaker of German (Krennmayr) go through all the errors and mark them up for possible L1 influence. After having calculated the proportion of errors that contained metaphor, we then calculated the proportion of those that were affected by L1 influence. Focusing on this smaller set of essays by German speakers of English allowed us to pilot our methodology for identifying errors and instances of L1 influence[4]. Both quantitative and qualitative findings are presented below. More details concerning the methodologies pertaining to our individual research questions are given where necessary.

## Results

In this section we present our findings with respect to each of the research questions listed above.

### In what ways does the amount of metaphor produced vary across CEFR Levels A2 to C2?

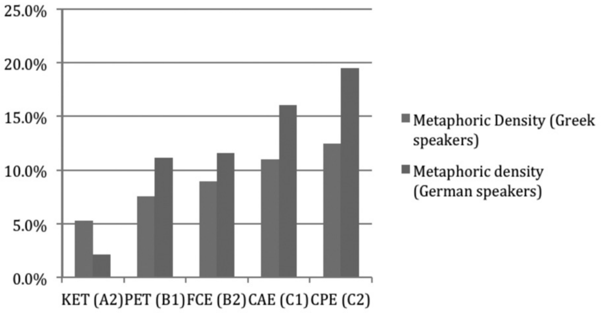
As we can see in Table 1 and Figure 2, in the essays written by the Greek-speaking learners, metaphoric density was found to start off fairly low but increased steadily across the levels, with statistically significant jumps from *KET* to *PET* (p<0.05) and from *FCE* to *CAE* (p<0.01). None of the other increases were significant. The overall trend in the data for the German-speaking learners was similar in that the only increases occurred between *KET* and *PET* (p<0.01) and between *FCE* and *CAE* (p<0.01) and between *CAE* and *CPE* (p<0.05). The main difference between the two data sets was that the German-speaking learners started off with a much lower level of metaphor at *KET* and that there was a statistically significant increase in metaphor from *CAE* to *CPE* in the essays written by the German speakers. The statistically significant increases from *KET* to *PET*, *FCE* to *CAE* and *CAE* to *CPE* are likely to be due to differences in the nature of the metaphor that the learners produce at these different levels, in response to the task demands, which in turn reflect the CEFR Can Do statements at that level; some examples are discussed below.

**Table 1: Metaphoric densities across levels in essays written by Greek-speaking and German-speaking learners**

| Level | No. of LUs (lexical units)<br>(Greek-speaking learners) | No. of LUs (lexical units)<br>(German-speaking learners) | No. of LUs containing metaphor<br>(Greek-speaking learners) | of LUs containing metaphor<br>(German-speaking learners) | Metaphoric density<br>(Greek-speaking learners) | Metaphoric density<br>(German-speaking learners) |
|---|---|---|---|---|---|---|
| KET (A2) | 744 | 800 | 43 | 17 | 5.8% | 2.1% |
| PET (B1) | 1,636 | 1,719 | 143 | 191 | 8.7% | 11.1% |
| FCE (B2) | 3,836 | 3,745 | 378 | 435 | 9.9% | 11.6% |
| CAE (C1) | 6,020 | 6,481 | 797 | 1,040 | 13.2% | 16.0% |
| CPE (C2) | 7,640 | 8,205 | 1,047 | 1,603 | 13.7% | 19.5% |

[4] Our long-term plan is to compare the effects of different L1 backgrounds on metaphor production in student writing.

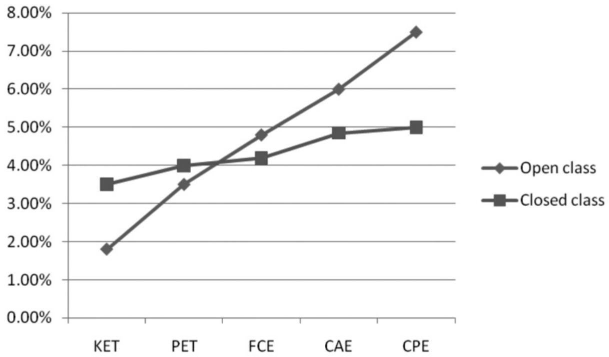
**Figure 2: Metaphoric densities across levels in essays written by Greek-speaking and German-speaking learners**



### In what ways does the use that learners make of open-class metaphorical items resemble or differ from that which they make of closed-class metaphorical items across the different CEFR levels?

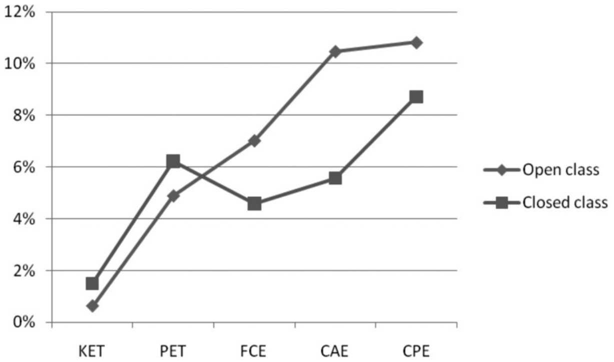
In addition to calculating how the overall metaphoric density changed across levels, the density based on whether the lexical units containing metaphor were open- or closed-class was also calculated. In the Greek data, the proportion of metaphoric open-class items was found to increase across levels with significant increases from *KET* to *PET* ($p<0.05$), *FCE* to *CAE* ($p<0.01$) and *CAE* to *CPE* ($p<0.01$), while closed-class items did not increase significantly, even across two levels. The proportion of metaphoric open-class items overtook the use of metaphoric closed-class items between the *PET* and *FCE* levels, as shown in Figure 3.

**Figure 3: Percentage of lexical units containing open and closed-class metaphor across levels in essays written by Greek-speaking learners**



In the German learners' essays, the proportion of metaphoric open-class items was found to increase across all levels with significant increases from *KET* to *PET* ($p<0.01$), *PET* to *FCE* ($p<0.05$) and *FCE* to *CAE* ($p<0.05$). The proportion of metaphoric closed-class items increased significantly from *KET* to *PET* ($p<0.01$) and from *CAE* to *CPE* ($p<0.01$). However it fell between *PET* and *FCE* and did not increase significantly from *FCE* to *CAE*. Again, the proportion of metaphoric open-class items overtook the use of metaphoric closed-class items between the *PET* and *FCE* levels.

**Figure 4: Percentage of lexical units containing open and closed-class metaphor across levels in essays written by German-speaking learners**



The fact that open-class metaphorically used items overtook closed items just before *FCE* (B2) in both groups of learners is interesting as it suggests that there is a qualitative change in the type of metaphor that the learners are starting to use at this level. This is likely to be a response to the tasks set, which generally require learners to state their opinions on certain issues and highlight their personal significance. This suggests that learners at *FCE* need to 'move up a gear' in their metaphor use; this may well be an experimental stage of language development during which they are particularly pushed to try out new metaphors. We return to this issue below.

### In what ways does the distribution of metaphor clusters vary across CEFR Levels A2 to C2?

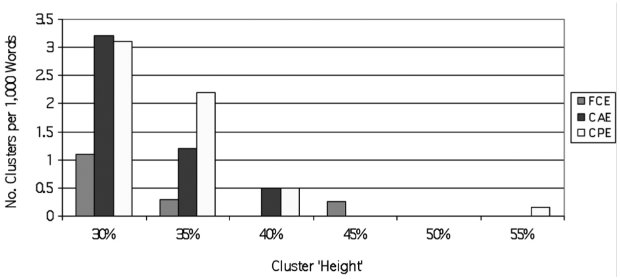
There was a marked increase in both the number of metaphor clusters and the density of these clusters at Level B2 of the CEFR in the essays written by the Greek learners, with no 30% density clusters appearing below that level:

**Table 2: Number of metaphor clusters appearing at each level in essays written by Greek-speaking learners**

| Level | Cluster count | No. of words | Clusters per 1,000 words |
|---|---|---|---|
| KET | 0 | 744 | 0.0 |
| PET | 0 | 1,637 | 0.0 |
| FCE | 6 | 3,838 | 1.6 |
| CAE | 29 | 6,020 | 4.8 |
| CPE | 46 | 7,688 | 6.0 |
| **Total** | **81** | **19,927** | **4.1** |

Both the number and the density of the clusters increased dramatically from then on, as we can see in Figure 5:

**Figure 5: Densities of metaphor clusters in essays written by Greek-speaking learners at FCE, CAE and CPE levels**

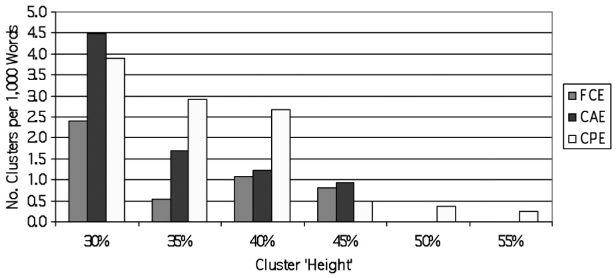


In the essays written by the German learners, a similar trend was observed, although clusters started to appear at *PET* (B1) level:

**Table 3: Number of metaphor clusters appearing at each level in essays written by German-speaking learners**

| Level | Cluster count | No. of words | Clusters per 1,000 words |
|---|---|---|---|
| KET | 0 | 800 | 0 |
| PET | 5 | 1,719 | 2.91 |
| FCE | 18 | 3,745 | 4.81 |
| CAE | 54 | 6,481 | 8.33 |
| CPE | 87 | 8,205 | 10.60 |
| **Total** | **164** | **20,950** | **7.83** |

**Figure 6: Densities of metaphor clusters in essays written by German-speaking learners at FCE, CAE and CPE levels**



With the exception of clusters of 30% and 45%, the cluster height increases as the students reach higher levels, a finding which corresponds to the Greek data in Figure 5 above. This suggests that students are becoming more confident in using metaphor at a more sustained rate, especially at Level C2 where clusters of 50% and 55% density are seen.

One of the clusters of 55% in a C2 level essay was as follows:

5 this appear to be a rather general advice that sounds vague. In order to turn it into a feasible concept, this process of learning should be applied to concrete situations. One issue, that is undoubtedly of importance, is humanity's attitude towards progress in science.
(German speaker of English, *CPE*: C2)

This can be compared to one of the clusters of 30% in the B2 level essays:

6 You can keep your body and your soul very fit. If you're a politician you will get a good image and get public interests.
(German speaker of English, *FCE*: B2)

These findings suggest that once learners have begun to use metaphor in clusters their ability to do so increases at a very fast rate. It is almost as if, at B2 level, learners start to get into a 'metaphorical mindset' which has a very positive effect on the quality of their L2 writing, as evidenced by the comparison of a 30% B2 cluster and a 55% C2 cluster above. In our data we observed considerable variation in the learners' tendency to use metaphor, which is in line with previous research showing that there are significant individual differences between learners in terms of their ability to comprehend and produce metaphor (Littlemore 2001). It would be useful if teachers could identify the skills involved in L2 metaphor production so as to foster this ability more widely among their learners.

**In what ways do the functions performed by the metaphor clusters vary across CEFR Levels A2 to C2 and how closely do these functions relate to the CEFR descriptors?**

In order to answer this question we conducted a manual search of all the essays at each level in order to identify the main functions, stylistic and phraseological features of the metaphors used. We were particularly interested in metaphorical features that had not appeared in our data at previous levels. We looked at metaphors that appeared in clusters as well as ones that did not. We hope to give a flavour of how the learners' use of metaphor develops over the five different levels in qualitative terms in the following discussion. Most importantly, we assess the ways in which the learners' use of metaphor helps them to achieve the Can Do statements at each level of the CEFR.

*Level A2*

The CEFR self-assessment grid for A2 level contains the following Can Do statement:

> I can write short, simple notes and messages relating to matters of immediate need. I can write a very simple personal letter, for example thanking someone for something.

It is difficult to see a clear role for metaphor at this level, except perhaps in the form of very dead metaphors within prepositions. This was confirmed by our data which showed that very little metaphor was used at this level (just above 5% for the Greek-speaking learners and less than 5% for the German-speaking learners). We can see that the metaphors in clusters at this level were mainly prepositions and fixed expressions, as we can see from this cluster taken from a German speaker's response:

7 They filmed us when we were studying and in our breaks. The programme will be shown on TV tomorrow at six.
(German speaker of English, *KET*: A2)

The lack of metaphor at this level clearly corresponds to the CEFR descriptor, as the learners are being asked to write clear notes containing factual information which will often involve dates and times.

*Level B1*

The CEFR self-assessment grid for B1 level contains the following Can Do statement:

> I can write simple connected text on topics which are familiar or of personal interest. I can write personal letters describing experiences and impressions.

Learners at this level are starting to use significantly more metaphor (particularly the German-speaking learners). In addition to using metaphorical prepositions, they are now beginning to use metaphor to present their own personal perspective, and to highlight the fact that they are providing their own perspective:

8 I was in your shoes last summer

9 Today I found time to give you some advice

(Greek speakers of English, *PET*: B1)

It is also at this level where we observe the first uses of personification metaphor:

10 these *companies* will realise the standards and regulations

(German speaker of English, *PET*: B1)

*Level B2*

The CEFR self-assessment grid for B2 level contains the following Can Do statement:

> I can write clear, detailed text on a wide range of subjects related to my interests. I can write an essay or report, passing on information or giving reasons in support of or against a particular point of view. I can write letters highlighting the personal significance of events and experiences.

At this level we have the beginning of persuasive language. Therefore one might expect an increase in the amount of metaphor used as this is one of the main functions of metaphor. Interestingly however, as we saw above, the overall amount of metaphor does not increase significantly at this level but perhaps more crucially, this is where the open-class metaphors start to take over in both the Greek and the German learner essays. At this level, some learners are able to use metaphor to provide reasons for and against their own points of view. In order to do this, they make more extensive use of personification metaphor:

11 They aren't really happy because money can't buy happiness.

12 mixed with other traffic, which takes not enough care of the bikers

(German speakers of English, *FCE*: B2)

Learners at this level are beginning to use metaphors with an evaluative function and it is here where we get the first what might be called 'creative' metaphors:

13 the only jewel that we have of transportations

(Greek speaker of English, *FCE*: B2)

Learners are beginning to use metaphor for dramatic effect in order to support their points of view:

14 They also consider it [the car] to be the bloodiest way of travelling.

(Greek speaker of English, *FCE*: B2)

They are also beginning to use metaphors that combine an evaluative function with a discourse organising function:

15 Bottom line is that. . .

16 I'd also like to point out that television can be a good company for lonely people

17 It is widely believed that

(Greek speakers of English, *FCE*: B2)

To sum up, although learners at this level are not using significantly greater amounts of metaphor than learners at level B1, they are using it to perform a much wider variety of functions and are making much more use of open-class metaphorical items.

*Level C1*

The CEFR self-assessment grid for C1 level contains the following Can Do statement:

> I can express myself in clear, well-structured text, expressing points of view at some length. I can write about complex subjects in a letter, an essay or a report, underlining what I consider to be the salient issues. I can select a style appropriate to the reader in mind.

One might expect an increased role for metaphor at this level as the learner needs to be able to express their points of view at length. Metaphor might also be involved in providing discourse coherence in essays that relate complex subjects. The word 'complex' might also be taken to include 'abstract' subjects, which would provide another role for metaphor, as metaphor is nearly always involved in the expression of abstract concepts. 'Underlining . . . salient issues' is also a form of evaluation that may involve metaphor. And metaphor might be involved in selecting a style that is appropriate for the reader. As for the issue of 'appropriate style', the use of metaphor constitutes a key feature of genre- and register-specific language (Deignan, Littlemore and Semino forthcoming, Semino 2008, Steen et al 2010).

In our data, the learners are starting to show clear evidence of an ability to use metaphors with the appropriate phraseology:

18 Even then Glenn did not rest on his laurels after reaching the top

(Greek speaker of English, *CAE*: C1)

In terms of functions, they are able to use metaphor to show relationships between their ideas and to reinforce their evaluations, as we can see in the following cluster:

19 On the one hand more women are taking part in *working life* than ever, but on the other hand leading positions are still occupied by male managers.

(German speaker of English, *CAE*: C1)

As well as using the conventional 'on the one hand' and 'on the other hand' to provide coherence, this learner also makes metaphorical use of the word 'occupied' to convey an image of possible stubbornness and unwillingness to move, on the part of the male managers.

Learners are able to use mixed metaphors, sometimes in clusters, in order to express abstract and complex issues:

20 where does this lead us? what's the prognosis for the future generations?

21 up to a certain point hopefully when we all realize that children are the reflection and the product of our lives
(Greek speakers of English, *CAE*: C1)

As we saw in the introductory section, mixed metaphors such as these are very common, particularly when a writer wants to write persuasively about difficult issues, or to get a particularly important point across (Kimmel 2010).

Learners at this level are also able to use metaphor to highlight salience and write emotively about topics that they feel strongly about, as we can see in this extract:

22 I believe this is a black date for Greek history
(Greek speaker of English, *CAE*: C1)

23 but you have still to struggle very hard, especially facing increasing recession in Europe. I hope that I won't fall behind a male colleague in the middle-management, where I work
(German speaker of English, *CAE*: C1)

This second cluster, which is used right at the end of this particular learner's essay, appears to serve a strong, evaluative, rounding-off function, relating the subject of the essay back to the learner's personal experience.

Some learners are able to use personification metaphors for persuasive or rhetorical effect:

24 the natural place for the [Parthenon] marbles to be was at their country at their home

25 his words were completely speaking inside my heart
(Greek speakers of English, *CAE*: C1)

Some learners at this level are starting to make use of direct metaphors (such as similes):

26 At that time his work was characterised as a candle in the wind
(Greek speaker of English, *CAE*: C1)

27 They still manage our country as they do the housework bring up the children and manage a little company called family
(German speaker of English, *CAE*: C1)

The learner in the latter example makes particularly sophisticated use of metaphor. The first use of 'manage' could be due to L1 influence, as the use of German 'managen' would sound appropriate in this context. The learner then turns this to their advantage and talks about 'managing a family' in the same way as they talk about 'managing the country'.

Some learners at this level are able to use metaphor to create dramatic contrasts. At times both halves of the dramatic contrast involve metaphor:

28 Once having the dream job, the nightmare starts.
(German speaker of English, *CAE*: C1)

At other times, dramatic contrasts are achieved by contrasting a literal meaning with a metaphorical one:

29 As for me, food may relief my hunger but his work is feeding my soul and spirit
(Greek speaker of English, *CAE*: C1)

To sum up, not only does the range of functions that learners are able to perform through metaphor expand considerably at this level, but they are starting to develop a strong sense of register. The metaphors are being deliberately used and manipulated (and at times played with) in order to achieve maximum rhetorical effect.

*Level C2*

The CEFR self-assessment grid for C2 level contains the following Can Do statement:

> I can write clear, smoothly-flowing text in an appropriate style. I can write complex letters, reports or articles which present a case with an effective logical structure which helps the recipient to notice and remember significant points. I can write summaries and reviews of professional or literary works.

One would expect here that an ability to use metaphor effectively is likely to contribute to a learner's ability to select an appropriate style and to highlight significance. Reviews of professional or literary works involve an ability to subtly ally oneself with, or distance oneself from, the work in question. Metaphor has been found to contribute to deictic positioning with respect to abstract concepts. For example, ideas that the author does not want to ally him or herself with are sometimes metaphorically construed as being further away ('that idea' as opposed to 'this idea') or in the past ('research suggested' as opposed to 'research suggests'). Metaphor is therefore likely to be involved in performing these subtle evaluative functions.

In our data, learners at this level are even more adept at using metaphors with appropriate phraseology and collocations:

30 Travelling makes you broaden your horizons

31 strengthen the bonds between nations
(Greek speakers of English, *CPE*: C2)

However, at times they are able to use metaphor with non-conventional, creative collocations to support their points of view:

32 In the midst of poverty and filth Mother Teresa has managed to create islands of hope where dignity is returned to those poor people who would otherwise despair out on the streets.

33 If we are conscious about the mistakes all our ancestors and former societies have done, we will not trudge into the same traps.
(German speakers of English, *CPE*: C2)

The expression 'trudge into the same traps' is not conventional in English but it is immediately comprehensible and the tr_ tr_ alliteration makes the expression particularly vivid, memorable and persuasive. By using this metaphor, the writer is able to present his or her opinion in very forceful terms.

Learners at this level are able to make creative use of direct metaphor to present their evaluations and points of view:

34 Our present values are not as firm as concrete. They can change again!

35 (. . .) your heath [health] will suffer when you reath [reach] a higher age. An old car doesn't run as smooth as a new one. This will sooner or later reduce your quality of life.

(German speakers of English, *CPE*: C2)

Personification metaphors are used in a more sophisticated way than at previous levels:

36 It has also given them a willing slave – the machine – which will work as many hours as required without demanding overtime or rest-time and without going on strike.

(Greek speaker of English, *CPE*: C2)

37 For the troubled *state coffers*, relief can be accomplished by structural changes in social security systems

(German speaker of English, *CPE*: C2)

In the first example, the personification of the machine forms the basis of an extended analogy. In the second example, the writer combines a personification/reification metaphor with a metonymy. This allows them to pack a large amount of information into a relatively short sentence leading to writing that has a more erudite and academic sound to it. It is related to the process of grammatical metaphor, which has been found to be an important feature of academic writing (Halliday 1985). In some cases, learners combine personification with more overt metaphors to add gravitas to their opinions:

38 It is the important task of *trade unions*, *companies* and politicians to try to make certain agreements which allow us to break through this vicious circle that is caused by prescriptions of the law.

39 In actual fact it should be the turn of the government now to take steps towards improvement.

(German speakers of English, *CPE*: C2)

Learners at this level are able to use metaphor, combined with metonymy to relate one part of their essay to another. In the following extract, the learner has been talking about politicians, but then they make metonymic use of the inclusive 'us' and 'our' to turn the reader's attention to more mundane, everyday events:

40 But let *us* look at the more obvious impacts on *our* own lives

(German speaker of English, *CPE*: C2)

Learners at this level are able to produce a high number of semi-coherent clusters, many of which contain peripheral response:

41 I firmly believe that putting one's life under the microscope severely affect the celebrity under investigation

(Greek speaker of English, *CPE*: C2)

42 A reaction one could have foreseen when looking back into history.

43 Inner values are certainly an asset for a person but in today's competitive environment self selling has become an important point.

(German speaker of English, *CPE*: C2)

Peripheral response is a phenomenon, first observed by Cameron and Low (2004) in which the metaphoricity of items that appear metaphorically dead is 'brought to life' by their proximity to metaphors from the same source domain occurring in the cluster. In the first example above, for instance, the term 'under investigation' appears more metaphorical than it might ordinarily appear because it is used in close proximity to the words 'under the microscope'. Peripheral response is common in both spoken and written English. It is always difficult to tell whether or not a writer has deliberately used metaphor in this way or whether they have done so subconsciously. However, the third example is likely to contain a degree of deliberateness as the writer then goes on to say:

44 We at IBM recommend our sales staff to wear clothes which match the style of their customers.

(German speaker of English, *CPE*: C2)

Learners at this level are also to convey sarcasm through metonymy:

45 Kohl ought to know very well about Germany's historical development since he had passed his exam in history.

(German speaker of English, *CPE*: C2)

46 We constantly hear proud announcements from *industry*

(Greek speaker of English, *CPE*: C2)

The second example is interesting as it combines a personification metaphor with metonymy, a phenomenon which we found to be common at higher levels.

Thus we have seen the increasing sophistication with which learners are able to use metaphor at each level. The functions that they are able to perform using metaphor map clearly on to the Can Do statements thus showing how the ability to manipulate metaphor effectively contributes to language development across the CEFR levels.

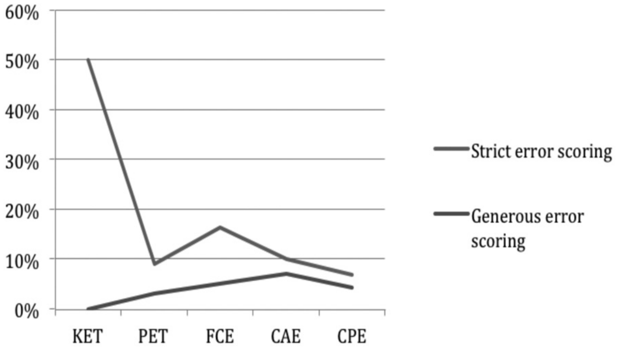
### To what extent do learners use metaphor 'incorrectly' and how is their use of metaphor influenced by their L1 background?

In order to answer the first part of this question, we took five essays at each level from our German-speakers' corpus and calculated the percentage of metaphors at each level that contained an error of some sort. As we saw above, we used both a strict error scoring procedure and a generous error scoring procedure. The findings are shown below:

**Table 4: Metaphors containing error at each level in essays written by German-speaking learners**

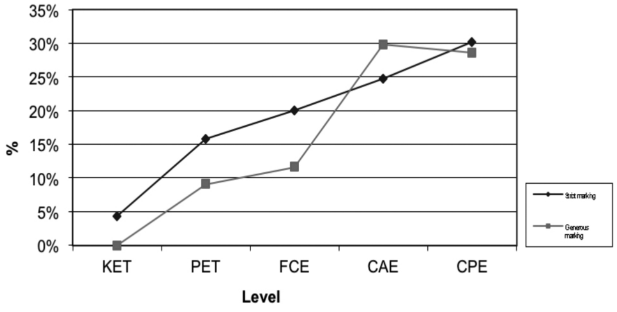
| Percentage of metaphors containing error | | | | | |
|---|---|---|---|---|---|
| **Level** | **Total metaphors** | **Total metaphors containing strict error** | **Total metaphors containing generous error** | **Strict error scoring** | **Generous error scoring** |
| KET | 2 | 1 | 0 | 50% | 0% |
| PET | 33 | 3 | 1 | 9% | 3% |
| FCE | 98 | 16 | 5 | 16% | 5% |
| CAE | 240 | 24 | 17 | 10% | 7% |
| CPE | 425 | 29 | 18 | 7% | 4% |

**Figure 7: Percentage of metaphors containing error at each level in essays written by German-speaking learners**



In Figure 7, we must first point out that the large drop in the strict error scoring line from *KET* to *PET* is not statistically significant as there were only in fact two metaphors used at this level. What is interesting here is the significant increase in metaphors containing error between *PET* and *FCE* under the strict scoring criteria (p<0.05). It trails off again towards *CPE*, though these decreases are not significant. What appears to be happening here is an experimental stage around *FCE* where learners try new things out and, as a result, they make more errors. This may well coincide with the fact that this is the level at which there are important qualitative changes in the use of metaphor. As discussed above, it is at this level where they start to use more open-class metaphor than closed-class metaphor. As they move through to the higher levels they start to use metaphor more correctly. However, if we compare their error rates with metaphor with their overall error rates, we can see that the percentage of errors involving metaphor actually increases in general terms (Figure 8).

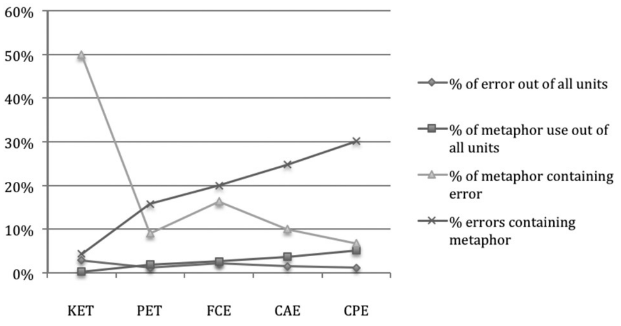
**Figure 8: Percentage of errors containing metaphor in essays written by German-speaking learners**



This is connected to the fact that the metaphor density increases steadily across levels, so errors involving metaphor make a greater contribution to the overall error count.

When we compare the trends of the general error rate with the metaphor error rate, we see that at *FCE* (B2) both error rates go up. At *CAE* (C1) both the metaphor error rate and general error start decreasing:

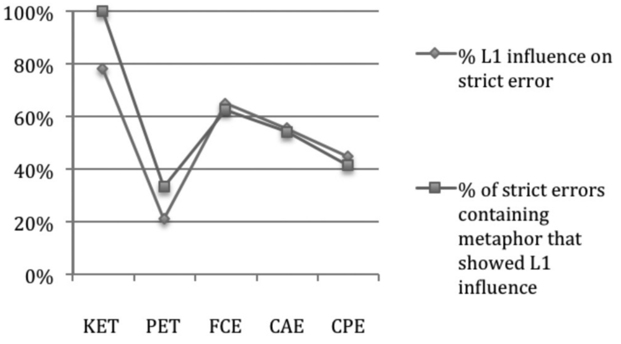
**Figure 9: Comparing the overall error rate, the percentage of metaphors use, the percentage of metaphors containing error, and errors containing metaphor in essays written by German-speaking learners**



This suggests that the rate of improvement for metaphor errors and the rate of improvement for other errors are correlated. This does not suggest, however, that metaphor is a phenomenon in language learning that does not need special attention. Note especially that the metaphor error rate is much higher than the overall error rate. Metaphor errors contribute to the overall error rate in a disproportionately large way compared to the amount of metaphor that is actually produced. This indicates that at any stage of learning, learners are more likely to make more errors when using metaphor than when using other types of language. This suggests that metaphor is something that teachers could usefully focus on throughout the learning process.

A native speaker of German (Krennmayr) then assessed whether any of the errors could be attributed to L1 influence. The results are shown in Figure 10:

**Figure 10: Percentage of L1 influence on errors and percentage of errors containing metaphor in essays written by German-speaking learners**



Beginning learners heavily rely on their native language. L1 influence decreases significantly from *KET* to *PET* (p<0.05), but then increases significantly (p<0.05) between *PET* and *FCE*. It weakens again as learners move up the CEFR levels but not significantly so. At the same time, the proportion of strict errors containing metaphor that show L1 influence follows the same pattern although only the *PET* to *FCE* increase is significant (p<0.05). The drop in level from *KET* to *PET* is not significant because of the very small number of cases of metaphor (N=1) at *KET* level. L1 influence on metaphor error starts to weaken gradually at C1 level but this decrease is not significant. This pattern again suggests that interesting things start to happen around *FCE* (B2) level. Learners make more errors at this level and L1 influence is more likely to be found here.

This L1 influence takes different forms. We have identified four types of L1-influenced errors in the use of metaphorically used words. The first type ('Type 1 errors') comprises errors that are not peculiar to metaphor in particular. An example is 'is everything running smooth' (German speaker of English, *FCE*: B2), which has to do with the difficulty German L1 speakers face in making a distinction between adjectival and adverbial forms.

The remaining error types are all metaphor related, albeit in different degrees. The clearest cases of metaphor errors ('Type 2 errors') are those of incorrect choice of a metaphorically used word, as illustrated in the examples below:

47 TV reports have wrapped their reports in dramatic pictures
(German speaker of English, *CAE*: C1)

48 the government has to force the production of bicycles
(German speaker of English, *FCE*: B2)

In both cases, error in metaphor use is likely to be due to L1 influence. For example, in the second excerpt, where the metaphor 'speed up' would have been a correct choice, the learner uses the inappropriate metaphor 'force' based on a transfer from the German verb 'forcieren.'

Learners may well choose the appropriate metaphorically used word but may not use it in its appropriate form due to L1 influence ('Type 3 errors'). Consider the following example:

49 this can cause a more sinister effect than nearly causing depressions
(German speaker of English, *FCE*: B2)

In English the basic meaning of 'depression' can be used in the plural form, whereas the metaphorical sense of a medical condition can only be used in the singular. In German, however, the metaphorical sense without a determiner is usually in the plural form.

The fourth metaphor error category ('Type 4 errors') comprises errors due to incorrect phraseology. Consider the following examples:

50 before end of next week
(German speaker of English, *CAE*: C1)

51 he started as nobody
(German speaker of English, *FCE*: B2)

52 famous people complain about having not enough private life
(German speaker of English, *CAE*: B2)

The use of 'nobody' without a determiner is generally used literally in English, whereas with a determiner ('a nobody') its use is always metaphorical. The wrong word order in 'having not enough private life' would feel slightly less wrong if 'life' were replaced by a concrete concept (e.g. 'to eat'), which would render 'having' non-metaphorical. The latter two metaphor examples in particular suggest the need for further research contrasting metaphorical and non-metaphorical uses, specifically in language learning contexts. We also need more detailed studies looking into the different types of errors listed in this section in order to develop specific guidelines for teachers as to which errors they need to address.

## Conclusion

At the beginning of this article, we outlined the two aims of the research, which were to identify features of metaphor that distinguish the different CEFR Levels A2-C2, and to provide descriptors relating to metaphor use that could be incorporated into the different CEFR descriptors for each level of writing for English.

We can summarise our main findings as follows:

- The proportion of metaphor used by language learners increases across the five CEFR levels studied.
- More open-class metaphors than closed-class metaphors are used from Level B2 onwards.
- Metaphor clusters start to appear at Levels B1 and B2.
- Metaphor is used to serve very different functions at each of the levels (see below).
- Rates of error involving metaphor are much higher than general rates of error across all levels of the CEFR.
- Rates of error involving metaphor and L1 transfer involving metaphor mirror general rates of error and L1 influence in that they peak at B2.

We would like to propose the following set of descriptors involving metaphor use for each level of the CEFR:

A2 Learners should be able to make accurate use of a limited range of metaphorical prepositions.

B1 In addition to the above, learners should be able to use a limited number of conventional metaphors, with appropriate phraseology in order to present their own perspective. They should also be able to make limited use of personification metaphors. They may be starting to use a small number of metaphor clusters.

B2 In addition to the above, learners should be able to make use of a limited number of conventional and creative open-class metaphors. They should be able to use metaphors for evaluative purposes and for dramatic effect and start to use them for discourse organising purposes. They should be starting to use personification metaphors more extensively. They should be starting to produce metaphor clusters, which may be coherent or contain mixed metaphors.

C1 In addition to the above, learners should be able to use direct, indirect and personification metaphors in clusters, with appropriate phraseology, for persuasive or rhetorical effect, to write emotively about topics that they feel strongly about, to show relationships between their ideas and to reinforce their evaluations. They may also use metaphor to create dramatic contrasts.

C2 In addition to the above, learners should be able to use metaphors with consistent appropriate phraseology and collocations, use non-conventional, creative collocations and make creative use of direct metaphor to present their evaluations. They should be able to produce a high number of semi-coherent clusters, possibly containing mixed metaphors and peripheral response. They may use personification metaphors as part of extended analogies and in combination with metonymy, and they may be able to convey sarcasm through metaphor and metonymy.

Based on the findings presented here, we would like to make a number of recommendations for ELT professionals. Firstly, textbook writers should consider introducing open-class metaphor and metaphor clusters at Level B2. Where possible, teachers should focus on expressivity rather than accuracy at Level B2 as this is where learners are switching from open to closed-class items and are starting to use metaphor in new ways which means that errors and L1 influence are particularly likely to occur at this level. Finally, teachers and syllabus designers should be aware of the fact that metaphor serves very different functions at different levels of the CEFR, and that it does not only appear in idioms. When Cambridge ESOL professionals are setting and marking written English at the different CEFR levels, it would be useful to take account of these important yet varied functions of metaphor at each level and to include at least some of the above descriptors in their marking criteria. More research is needed to investigate the more subtle phenomenon of metonymy and it would be very interesting to explore the roles played by both metaphor and metonymy in spoken language production. While we would expect some overlap, there are important differences between written and spoken language that lead us to hypothesise a different set of results. These differences are twofold. For understandable reasons, written and spoken descriptors of proficiency are categorised differently, and more importantly, the functions of spoken versus written language are not the same.

## References and further reading

Alejo, R (2010) Making sense of phrasal verbs: a cognitive linguistics account, in Littlemore, J and Juchem-Grundmann, C (Eds) Applied cognitive linguistics in second language learning and teaching, *AILA Review* 23, 50–71.

Biernacka, E (forthcoming) *The role of metonymy in political discourse*, unpublished PhD thesis, The Open University, Milton Keynes, UK.

Cameron, L and Low, G D (2004) Figurative Variation in Episodes of Educational Talk and Text, *European Journal of English Studies* 8 (3): 355–374.

Cameron, L and Stelma, J (2004) Metaphor Clusters in Discourse, *Journal of Applied Linguistics* 1 (2), 107–136.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.

Deignan, A (2005) *Metaphor and Corpus Linguistics*, Amsterdam/Philadelphia: John Benjamins.

Deignan, A, Littlemore, J and Semino, E (forthcoming) *Metaphor in Discourse Communities*, Cambridge: Cambridge University Press.

Halliday, M A K (1985) *An Introduction to Functional Grammar*, London: Arnold.

Kellerman, E (1987) *Aspects of Transferability in Second Language Acquisition. A Selection of Related Papers*, PhD dissertation, Nijmegen: University of Nijmegen Press.

Kellerman, E (1987a) An Eye for an 'Eye', in Kellerman, E, *Aspects of Transferability in Second Language Acquisition. A Selection of Related Papers*, 154–177.

Kellerman, E (1987b) Towards a Characterisation of the Strategy of Transfer in Second Language Learning, in Kellerman, E, *Aspects of Transferability in Second Language Acquisition. A Selection of Related Papers*, 89–124.

Kimmel, M (2010) Why we mix metaphors (and mix them well): Discourse coherence, conceptual metaphor, and beyond, *Journal of Pragmatics*, 42, 97–115.

Littlemore, J (2001) Metaphoric competence: a possible language learning strength of learners with a holistic cognitive style? *TESOL Quarterly* 35 (3), 459–491.

Littlemore, J and Low, G (2006a) Metaphoric competence and communicative language ability, *Applied Linguistics* 27 (2), 268–294.

Littlemore, J and Low, G (2006b) *Figurative Thinking and Foreign Language Learning*, Basingstoke/New York: Palgrave Macmillan.

MacArthur, F and Littlemore, J (forthcoming) On the repetition of words with the potential for metaphoric extension in conversations between native and non-native speakers of English, *Metaphor and the Social World.*

Müller, C (2008) *Metaphors, Dead and Alive, Sleeping and Waking. A Dynamic View*, Chicago: University of Chicago Press.

Pragglejaz Group (2007) MIP: A method for identifying metaphorically used words in discourse, *Metaphor and Symbol* 22 (1), 1–39.

Semino, E (2008) *Metaphor in Discourse*, Cambridge: Cambridge University Press.

Sinclair, J (1991) *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Steen, G, Dorst, A G, Herrmann, J B, Kaal, A, Krennmayr, T and Pasma, T (2010) *A Method for Linguistic Metaphor Identification*, Amsterdam: John Benjamins Publishing Company.

# The attitudes of teachers and students towards a PET-based curriculum at a Japanese university

**JUN NAGAO, TORU TADAKI, MAKIKO TAKEDA AND PAUL WICKING,** MEIJO UNIVERSITY, NAGOYA, JAPAN

## Introduction

As the Common European Framework of Reference for Languages (CEFR; Council of Europe 2001) extends its influence worldwide, curriculum developers will be looking for ways to incorporate this framework into their courses. The Cambridge English exam suite is one option, aligned to the CEFR levels, around which an English as a Foreign Language (EFL) program can be structured. This article reports on a study that looks at a university in Japan which did exactly this. While the university program incorporates KET, PET and FCE- based curricula, the focus of this research is the PET exam. The attitudes and perceptions of 31 first-year Japanese university students towards the PET are documented and analysed, as well as their performance in the test over the course of one academic year. Likewise, teacher beliefs and opinions of the PET are presented and examined.

Japan has achieved somewhat of a reputation internationally as a nation of test-takers. There is very high prestige placed on the ability to score well on tests, and especially so on tests of English ability. Hundreds of thousands of Japanese students take general English tests every year, most notably TOEIC[1] (Test of English for International Communication) and EIKEN[2] (Test in Practical English Proficiency). However, despite being internationally recognised and being widely taken elsewhere, the uptake and performance of students taking Cambridge ESOL's Preliminary English Test (PET) has been relatively low within Japan to date (for example, the total pass rate at B1 Level for PET in 2009 was 26.3%, the second-lowest pass rate worldwide, see Cambridge ESOL 2011a).

Previous studies of the washback effect of tests have had mixed findings. Watanabe (2000) found a significant amount of negative washback on teachers, as did Shohamy, Donitsa-Schmidt and Ferman (1996). Some studies have noted that teachers felt that they were 'teaching to the test' (Alderson and Hamp-Lyons 1996, Read and Hayes 2003), while other researchers have found that teachers often worry that students will be overly focused on passing the test rather than the goal of actually learning a language (Buck 1988, Raimes 1990, Shohamy 1992). Torikai (2010) notes that despite the increasing importance of the TOEIC test for businesses in Japan, there are some serious limitations as to what conclusions can be drawn from individual results.

Washback is also found to have some positive effects. Lewthwaite (2007) documents the positive washback of the IELTS writing tasks on both students and teachers at a university in the Gulf. Both teachers and students alike found it to be a reasonable and appropriate communicative test of writing performance, and relevant to real-world skills.

Concerning the area of Cambridge ESOL exam washback in the Japanese classroom, it appears that only a single (unpublished) study has been completed. Harwood (2007) looked at the washback of the Key English Test (KET) on a Japanese high school. While there were both negative and positive aspects of washback, Harwood found both teachers' and students' perceptions about teaching and learning toward the KET were heterogeneous and often contradictory. The test preparation textbook used in this school, *Objective KET* (Capel and Sharp 2005) was felt to be appropriate. However, teachers believed the KET was 'Eurocentric' and needed more topic areas related to the Asian context. Informal comments from teachers at the university beforehand suggested that this study would find similar results.

## Research questions

The main objective of this research was to examine the attitudes and perceptions of teachers and students at a Japanese university toward PET and a PET-based curriculum. As such, there were two ultimate goals. The first was to chart the changing perceptions of Japanese university students toward the PET over the course of one academic year. In particular, attention was paid to the needs of those students and to the extent in which a program of study based around the PET meets those needs. The second was to assess teacher attitudes towards the PET and the program designed around it. Therefore, this study addresses the following four research questions:

1. Is PET an appropriate target for the needs of students at this university? If not, how is it being or should it be changed to fit this context?
2. How do students themselves feel about the format and content of the test?
3. Are teachers positively or negatively disposed towards the PET exam, and what are the reasons for this?
4. How do teachers feel about the program and the materials provided?

Cambridge ESOL's Preliminary English Test corresponds to Level B1 of the Common European Framework of Reference for Languages (CEFR). There are three sections. The Reading and Writing paper is worth 50% of the final grade, and candidates are allowed 90 minutes to complete it. The Listening paper is worth 25%, and lasts about 30 minutes. The Speaking test is also worth 25%, and candidates take the test in pairs. The focus of the exam is everyday written

[1] See http://ets.org/toeic

[2] See http://stepeiken.org

and spoken communication; it is useful for students studying English for work and travel purposes and it provides a step towards higher level qualifications (Cambridge ESOL 2009). The *PET Handbook for Teachers* states 'candidates who are successful in PET should be able to communicate satisfactorily in most everyday situations with both native and non-native speakers of English' (ibid. 6).

## Research context

This research was conducted at a major university in the Chubu area of Japan. The university has eight faculties; namely: agriculture, business management, economics, human studies, urban science, pharmacy, science and technology, and law, of which the first six have joined the liberal arts educational program.  There are no students who major in English. The number of students taking compulsory English classes in the liberal arts educational program is about 2,700. There are five levels in the English program: basic, elementary, pre-intermediate, intermediate and advanced. The entire program is based on the CEFR. The program is officially called 'the Liberal Arts English Program' (hereafter LAEP).

The five levels of the LAEP correspond to the three broad levels of CEFR (A, B, C) and the Cambridge English exam suite. The advanced course has been developed based on the B2 level of the CEFR, while the intermediate and the pre-intermediate courses have been designed according to the B1 level in the CEFR scale. The content of the elementary and the basic courses has been developed according to the A2 level in the CEFR scale. The corresponding Cambridge English exam levels are shown in Table 1.

**Table 1: Corresponding levels between the LAEP, CEFR and Cambridge ESOL tests**

| LAEP | CEFR | Cambridge ESOL exam suite |
|---|---|---|
| Advanced | B2 | First Certificate in English (FCE) |
| Intermediate | B1 | Preliminary English Test (PET) |
| Pre-intermediate | | |
| Elementary | A2 | Key English Test (KET) |
| Basic | | |

The 412 classes in the LAEP are taught by over 50 teachers. Almost all of these teachers are employed part-time on a contract basis. The average class size is between 20 and 30 students. Students are taught by a native English speaker (NES) teacher once a week, as well as a non-native English speaker (NNES) teacher once a week. This means that for one student, there are two 90-minute lessons a week. NES teachers focus on the productive skills (speaking and writing) while NNES teachers focus on the receptive skills (listening and reading). The NES and NNES teachers use the same textbooks in their classes, and the medium of instruction is both English and Japanese. At the pre-intermediate level, these are *Insight into PET* (Naylor and Hagger 2004) and *English Vocabulary in Use: Pre-intermediate and intermediate* (Redman 2003). At the intermediate level, *PET Masterclass* (Capel and Nixon 2003) and *English Collocations in Use* are used (McCarthy and O'Dell 2005).

The content and format of the end of semester exams are almost identical to the Cambridge English KET, PET and FCE tests. All four language skills – speaking, writing, listening and reading – are assessed with these semester final exams, which is unusual for a Japanese university. In the context of English education in Japan, it is quite rare to assess any skill or knowledge other than grammar and translation (Shizuka 2002, Wakabayashi and Negishi 1993). The teachers in this program are required to teach their students towards these exams. In this sense, it is an exam-oriented program. However, as these exams are designed to be a test of communicative English ability, it could also be said that the teachers are required to teach their students to enable them to actively communicate: to speak, to listen to understand, to read and write in English in its real sense.

## Data collection

In order to examine the attitudes and opinions of teachers and students towards the program, questionnaires and interviews were conducted over the course of one academic year. In the first semester, two different questionnaires were given to 31 first-year Japanese university students. The respondents, of whom 29 were intermediate students and 2 were pre-intermediate, were drawn from 11 different classes. The first questionnaire was conducted in the second week of the first semester in April 2010 to find out students' English learning background and their needs. The questionnaire contained 10 questions using a five-point Likert scale and nine open-ended questions to elaborate on their answers. The second questionnaire was given immediately after taking their first PET exam on May 22nd 2010, to record their first impression of the PET. This was an official PET exam, separate from their English course, held by a licensed testing centre. The second questionnaire consisted of 17 questions using a five-point Likert scale and three open-ended questions. For each Likert scale question, students were asked to write the reasons for their answers.

In addition to the first two questionnaires, a third questionnaire was given in the second semester to the same students after taking a second official PET exam on December 4th 2010, to find out how those students felt about PET and how their perceptions towards PET had changed through following the PET-based curriculum. One student was absent, therefore 30 students took the second PET and answered the third questionnaire. In order to see the changes, if any, most of the questions in the third questionnaire were the same as those of the second questionnaire to compare the answers between them. The third questionnaire therefore comprised 16 questions using a five-point Likert scale and four open-ended questions. Students were also asked to explain their answers for most of the Likert scale questions.

The attitudes of teachers are of vital significance for any study of washback. In her review of the literature concerning washback, Spratt notes 'the teacher is constantly mentioned as playing a pivotal role in determining whether washback occurs, how and to what degree' (Spratt 2005:21). Data was gathered from teachers by means of a questionnaire survey and an interview. The survey respondents were 21 part-time teachers working in the LAEP. The questionnaire had 15 questions, and participants were asked to rate their answers on a Likert scale from one to five. There was also a free response section, where participants could elaborate on their answers in greater detail. Responses were completely anonymous.

Participants in the interview, all volunteers, were four part-time teachers in the LAEP: two NES and two NNES. They were informed about the focus of the study, and were free to leave the study at any time. Each participant chose a pseudonym. They each have professional TEFL qualifications, and from 6-15 years of ELT experience. We focused on individuals who had spent some time in the profession, and so had a depth of experience from which to evaluate the PET. A semi-structured interview lasting between 30-40 minutes was conducted with each participant. The interviews were transcribed for analysis, and transcriptions were sent back to the participants for revisions and comments.

The results of the study are presented below.

### Student questionnaire results

*The students' needs for studying English*

According to the background questionnaire (Q.11, see Table 2), the reasons of the students for studying English are mainly extrinsic – they want to study English for practical reasons. For example, a third of the students (11) thought that they need to study English because English is useful or necessary in their future life. Almost the same number of students (10) thought that they want to study English to be able to communicate with foreigners. Seven students hoped that they would work or study overseas and six students thought they needed English to travel abroad. Three students wanted some English related qualifications.

Intrinsic motivation was not altogether absent. Six students wanted to study English simply because they like English. Three wrote that they want to be able to use English and two thought that being able to speak English is cool. Two students wanted to understand and enjoy English musicals, songs and films.

**Table 2: 'Why do you want to study English?'**

| Reasons for studying English | No. of students |
|---|---|
| English is useful/necessary in the future. | 11 |
| I want to communicate/speak with foreigners. | 10 |
| I want to work/study overseas. | 7 |
| I want to travel abroad. | 6 |
| I like English/enjoy learning English. | 6 |
| I want to be able to use English | 3 |
| I want some qualifications. | 3 |
| Being able to speak English is cool. | 2 |
| I want to be able to understand musicals, songs, movies.. | 2 |
| Others | 6 |

As far as the needs of the students for studying English are concerned, the data shows they want to be equipped with practical English abilities. This coincides with the purpose of the Cambridge English tests, which, according to Cambridge ESOL's website: 'give you the language skills you need to succeed in an English-speaking environment' (Cambridge ESOL 2011b).

*The students' expectations of university English education*

The majority of the students expect to use practical English at university (Background questionnaire, Q.18; see Figure 1). They want to be able to use English rather than study about English. This strong tendency in their expectations for a university English program might spring from their prior experience at junior and senior high school, where they were mainly taught English through grammar-translation (Post PET questionnaire, Q.5 and 7; see Figure 2).

**Figure 1: 'What do you expect to study in English class at university?'**

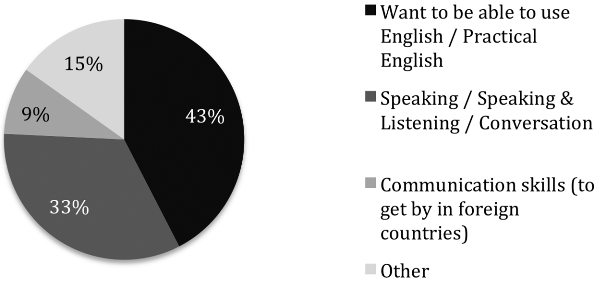
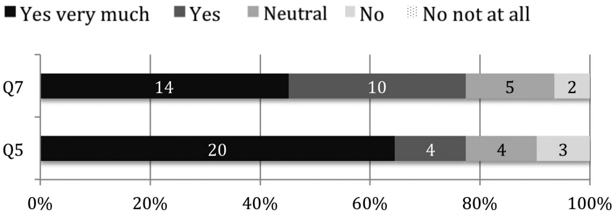


**Figure 2: 'The lessons at junior high school (Q5), high school (Q7) were given with grammar translation method'**

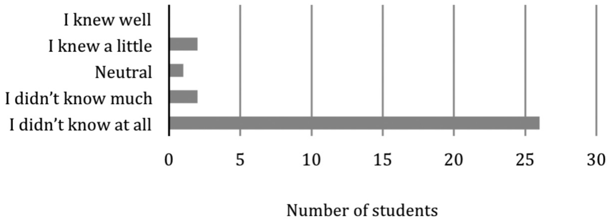


Students generally want to study English at university in order to be able to use English; specifically, to communicate with people from other cultures or to do things in English. These expectations also coincide with the purposes of the test given by Cambridge ESOL, which states on their website: 'You may be thinking of studying abroad or working in another country to fast-track your career. Either way, a Cambridge English certificate will take you where you want to go.' (Cambridge ESOL: 2011b). PET is aligned to the Common European Framework of Reference for Languages, which answers the students' expectations for achieving a globally recognised standard of English.

*The students' familiarity with PET*

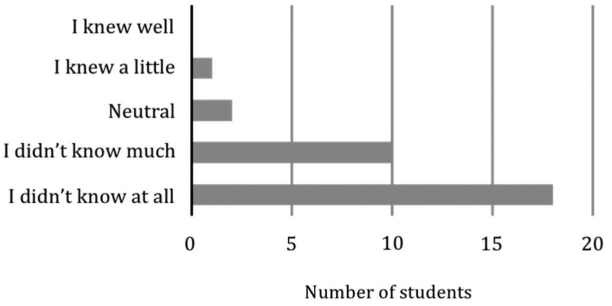
The data from the student questionnaire revealed that the vast majority of the students (84%) were not familiar with Cambridge ESOL exams before entering the university (see Figure 3). Evidently, it was up to the teachers to decide whether or not to inform the students of the benefits of taking PET, and the theoretical reasons behind adopting CEFR as the framework for the English program.

**Figure 3: 'I knew about Cambridge ESOL Exams before coming to [this university]'**



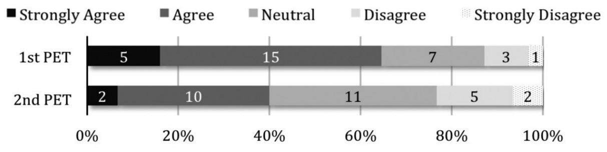
Most students were new to the testing format used for PET (see Figure 4), which is not surprising considering that they had never taken PET before. During the academic year, the LAEP curriculum gradually introduced students to the testing format.

**Figure 4: 'I knew the format of the PET'**



Although over half of the students said that they actually enjoyed taking PET (see Figure 5), this number decreased in the second questionnaire in December. Overall, the comments from the students suggest that how well they felt they did on the second exam influenced their answer for this question.

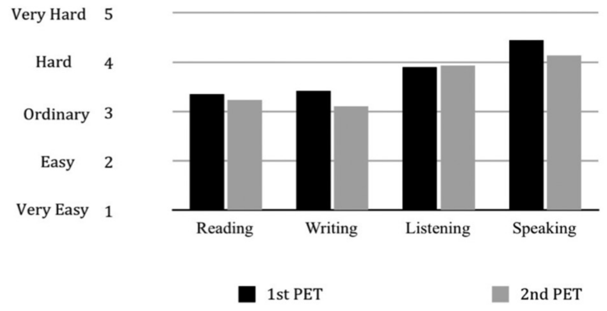
**Figure 5: 'Were you able to enjoy taking the PET?'**



*The students' impression of the components of PET*

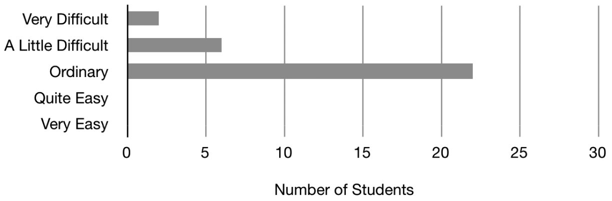
In both first and second questionnaires, the students were asked how they felt about each section in the PET. Figure 6 indicates the average scores of the students' answers.

**Figure 6: The students' impression of each section of PET**



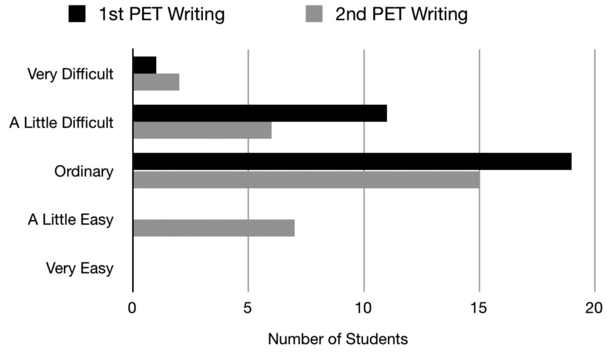
In the first questionnaire, for the difficulty of the reading section, 73% of the students responded 'ordinary,' 20% responded 'difficult,' and 7% responded 'very difficult' (see Figure 7). Among the students who responded 'ordinary,' 27% mentioned that the reading section of PET was similar to the 'Center Test' (a standardized university entrance exam) or readings they did in high school. Some students who responded 'a little difficult' or 'very difficult' said that it was difficult because they encountered new words. Others said the vocabulary used in PET was not difficult, but they were not sure how to read the texts. One student said that she understood the text but could not answer the questions, and another said that he could not read the text fast enough.

**Figure 7: 'How difficult was the reading section?'**



As for the writing section, when the first and second questionnaire results are compared, there is a slight decrease, indicating a number of students found the writing section in the second PET less difficult (see Figure 8). Over one-third of the students who responded 'quite easy' or 'ordinary' in the second questionnaire mentioned that practising writing in class or for homework helped them better prepare for the writing section of the PET.

**Figure 8: Level of perceived difficulty of the writing section**

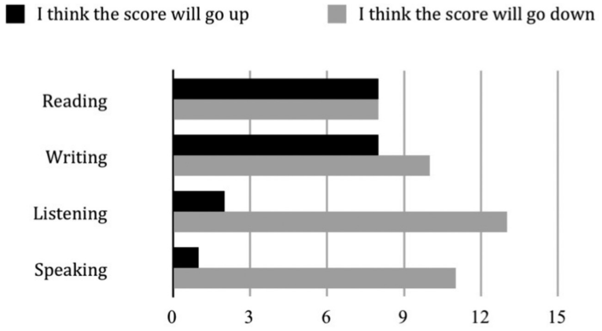


For the listening section in both questionnaires, almost half of the students who found the listening section 'a little difficult' or 'very difficult' felt that conversations and dialogues in the listening section were too fast.

The speaking section was found to be the hardest among all the sections in PET both times, although there was a slight decrease in the second questionnaire. There were comments such as 'I am not used to speaking English,' 'I couldn't find appropriate words to use and stopped talking,' and 'Sorry about causing a trouble when I went blank and froze.' Seeing these sorts of comments in both questionnaires suggests that more often than not, many students were not used to using English orally. Having one 90-minute communication class a week with 20 to 30 students was generally not enough to help them feel comfortable and confident about speaking English.

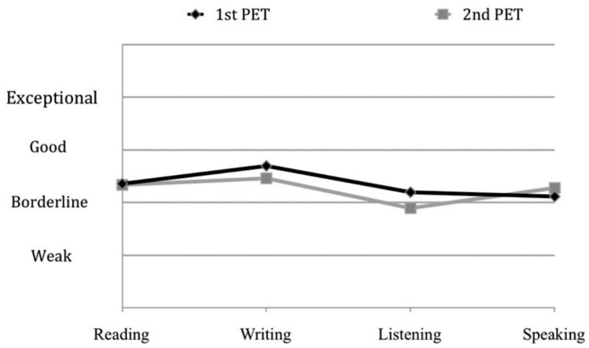
In response to the question in the second questionnaire that asked, 'Compared to the PET in May, which section do you think will show a change in score?' about half of the students thought that their scores for the reading section and writing section would go up (see Figure 9). However, for the listening and speaking sections, the majority expected that their scores would go down.

**Figure 9: 'Which section do you think will show a change in score?'**



However, students' actual scores showed an increase in the speaking section of the second PET, while the scores for the other sections slightly decreased as shown in figure 10.

**Figure 10: Average student PET scores in May and December**



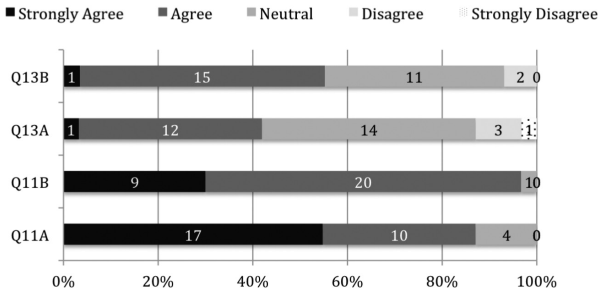
*Suitability of PET in real-life contexts*

The students thought that the PET is suitable for their English use in real-life situations (see Figure 11). Their responses to Q11 and Q13 of the post PET questionnaires clearly show this tendency. In their view, PET is not merely a test for language knowledge but for the real use of the language. The questions are as follows:

Q11: Skills for passing PET will be useful when I use English in a real-life context in the future.

Q13: The content of the reading section is related to real-life situations.

**Figure 11: Q11 & Q13 of the post PET questionnaires (A=first questionnaire, B=second questionnaire)**



The students believed the PET assesses useful skills in real-life situations mainly because PET includes a speaking test section, where they are assessed on their ability to express their opinions in English, which they thought would be a useful and necessary skill in their future life. Usually they do not know any other test which assesses four language skills. It is likely that PET is the only exam to their knowledge which tests speaking and writing ability at the same time. For them, productive skills, which are writing and speaking, are crucial in real-life situations, as shown in Tables 3 and 4.

**Table 3: Reasons for the response to Q11 (first post PET questionnaire)**

| Reason | No. of responses |
|---|---|
| Speaking and expressing my opinion are necessary skills | 10 |
| PET assesses all four skill areas | 3 |
| PET is widely recognized and good for boosting career prospects | 1 |
| PET covers a range of language contexts | 1 |
| Speaking & listening sections are useful | 3 |
| PET is relevant for travelling & living in foreign countries | 2 |
| PET has clear targets so I can get motivated | 1 |
| PET is useful for reading English books | 1 |
| PET is useful for boosting practical English ability | 3 |
| Other comments | 4 |
| No comment | 3 |

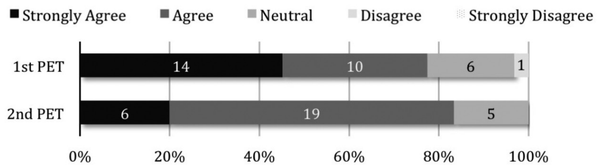
**Table 4: Reasons for the response to Q11 (second post PET questionnaire)**

| Reason | No. of responses |
|---|---|
| Speaking and expressing my opinion are necessary skills | 7 |
| PET assesses all four skill areas | 3 |
| PET is widely recognized and good for boosting career prospects | 3 |
| The level of PET is appropriate | 2 |
| Listening and writing sections are useful | 1 |
| PET covers a range of language contexts | 2 |
| Other comments | 4 |
| No comment | 8 |

*Students' perceptions towards PET*

The questionnaire results show that over 75% of the students in both questionnaires felt that PET is a test that accurately assesses their English proficiency.

**Figure 12: 'PET is a test that accurately assesses my English proficiency'**

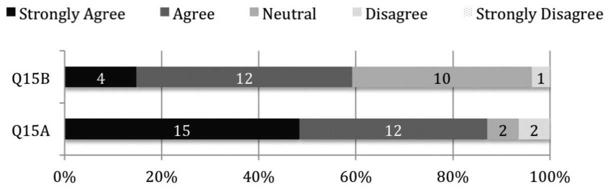


The most common reason that the students provided was that PET assesses all four skills, while other tests do not assess their speaking skills. The only student who responded 'disagree' in the first questionnaire said that 'there are various tests, and the results of various tests will show real English ability.' The reasons provided by students who responded 'neutral' in the first questionnaire come from their uncertainty towards their test results and the testing format because it was their first time taking PET. One student commented that 'the speaking test was conducted with a partner as opposed to one-on-one,' which suggests that the rationale behind the testing format are not always clear to the students. On the other hand, in the second questionnaire, there was no one who responded 'disagree'. Those who chose 'neutral' provided three comments which were:

'I felt that it was different from English I normally study,' 'I don't really know,' and 'There are various sections (i.e. speaking and writing).'

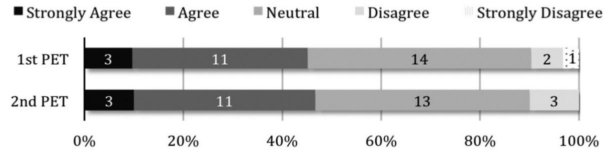
If we look closer at Q15 (Figure 13) of the first and second post PET questionnaires, the students responded to this question differently. In the first questionnaire (Q15A), they apparently appreciated the speaking test, presumably because that was their first opportunity to take a speaking test. Without any prior experience of taking a speaking test, it might be difficult to judge the appropriateness of that section. In the second post PET questionnaire (Q15B), however, the number of students who chose 'neutral' has increased from two to ten and the number of students who chose 'strongly agree' has decreased from fifteen to four. The precise reason for this is unclear, but it seems as if some thought their performance in the speaking test was not a true reflection of their speaking proficiency. Even so, in the second questionnaire, more than half the students thought that the speaking section of PET did accurately measure their speaking ability.

**Figure 13: 'The speaking test of PET accurately measures my speaking ability'**



In response to the statement, 'Topics in the PET are appropriate for people like me who live in Japan,' almost half of the students agreed or strongly agreed, while the other half were neutral or disagreed (Figure 14). The ratio for the students' perceptions on this issue did not change over time to a large extent.

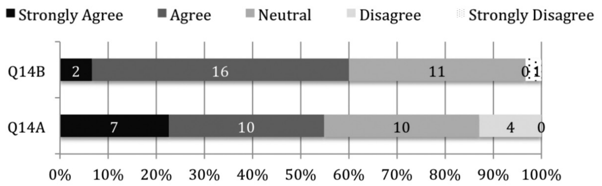
**Figure 14: 'Topics in the PET are appropriate for people like me who live in Japan'**



Students who 'strongly agreed' or 'agreed' commented that the topics are 'common all over the world' and 'useful when communicating with people overseas'. On the other hand, students who answered 'neutral' commented that they wanted the PET to include 'topics that are useful in daily life,' and that they feel 'some topics are only based in foreign countries.' Overall, the comments for this statement indicate that some students are rather well-disposed towards the topics in PET, while some wished that the topics would be more relevant to their lives in Japan.

When they were asked if the Cambridge exams were suitable for Japanese students, 60% of the students answered 'yes' (see Figure 15). The main reason for this response was that the Cambridge exams focus on four language skills. In other words, they appreciated PET because it assesses speaking ability. No one referred to cultural issues.

**Figure 15: 'Cambridge exams are suitable for Japanese students'**



*Students' motivation*

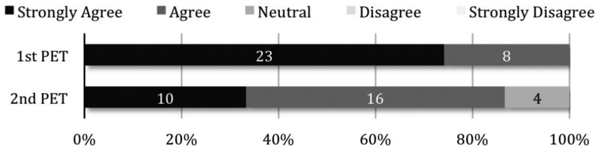
In the second questionnaire, the students were asked, 'Do you think your score will be better than last time?' 70% of the students answered 'no' and 30% answered 'yes'. 52% of the students who said 'no' said that their score would be worse because they did not study enough, 38% said the exam itself would be the cause (for example, it is just too difficult), and 10% gave other reasons such as their physical condition on the exam day.

Among the 30% of the students who chose 'yes', 50% said that their score would go up because they were more accustomed to PET than the last time they took it, 30% said that it was because they studied, and 20% gave their impression on how well they did on the exam as a reason.

This result suggests the students need constant encouragement or re-enforcement to continue studying. Also, they need to set up more precise goals besides simply passing PET. If passing PET were their ultimate goal, their motivation would drop after they accomplished it.

On the whole, as figure 16 shows, students felt more motivated to continue studying English after taking PET.

**Figure 16: 'After taking PET, I feel more motivated to study English'**

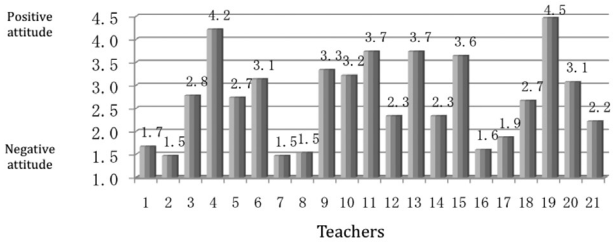


In the first questionnaire, to the statement "After taking PET, I feel more motivated to study English", the common reasons the students gave for choosing 'agree' or 'strongly agree' were that they didn't perform as well as they hoped on the exam (especially the speaking section) and they wanted to do better next time, or that they wanted to improve their English in general. Similar reasons were found in the second questionnaire. Many students mentioned that they were not satisfied with their English level. Some of the students said that they were 'shocked' to find out how poor their English was. The reason that the rate for choosing 'strongly agree' dropped is not clear from the comments. However, it seems safe to assume that they had a clearer goal of passing PET after taking the first PET, and once they finished taking the second PET, that goal was lost. Also, it may have been hard for those who passed PET in May to keep themselves motivated to study for PET in December.

### Teacher questionnaire results

Figure 17 shows the general tendency of each of the 21 teachers who responded to the survey, out of 50 teachers who were invited to respond. The figures are averages of individual teachers, where 5.0 is given to 'strongly agree', 4.0 to 'agree', 3.0 to 'neutral', 2.0 to 'disagree' and 1.0 to 'strongly disagree'. In general, a 'strongly agree' response indicates a strongly positive attitude, while a 'strongly disagree' response indicates a strongly negative attitude. (However, for questions 3 and 5, the figures are given in reverse; so 'strongly disagree' is indicative of a strongly positive attitude.) The figure indicates, then, that a teacher responds positively in general if the numerical value is above 3.0. There are six teachers whose values are less than 2.0 and this indicates that they have a strongly negative attitude towards PET and the program. On the other hand, a strongly positive attitude seems to be held by two teachers whose values are more than 4.0.

**Figure 17: Teachers' attitudes towards PET and a program designed around PET**



Generally, it seems that teachers' attitudes towards a course based on the PET are slightly more negative than positive. This is consistent with other studies of exam washback, which note the negative feelings that exam preparation generates with teachers (Alderson and Hamp Lyons 1996, Shohamy et al. 1996). While there were some questions which generated strong opinions from individual teachers, overall, the results indicated that teachers have ambivalent opinions of the PET course. Many questions generated answers within the 2.5~3.5 range, which indicates neutrality.

The most strongly negative response came from statement nine: 'The PET test is equally applicable in any cultural context', to which 66% of respondents either disagreed or strongly disagreed. Another notably negative response came from the eighth statement: 'Based on my experience with the PET exam, I would recommend other universities in Japan make use of the PET in their general English courses'; 38% strongly disagreed, 10% disagreed, 25% agreed and there were no strongly agree responses. It also appears that teachers believe they have a good grasp of the PET course and understand the purpose of it. For question three, 'I can easily understand this course and the aims it is trying to achieve,' 76% of answers indicated 'agree' or 'strongly agree'.

In the free response section of the survey, the textbooks used in the program came in for severe criticism. (These textbooks were not in any way endorsed by Cambridge ESOL.) One teacher wrote 'What I have the most trouble dealing with are the textbooks. In my opinion, they are not suitable for EFL students.' Many teachers felt this way, with nine responses of 'strongly disagree' to the statement 'The PET based textbooks work well with my students.' For a discussion of the complex role that textbooks play in test washback, see Hamp-Lyons (1998).

The European cultural slant was also criticised, with one respondent commenting 'I don't think that teaching only British English is appropriate for our students.' Some teachers also felt that the content of the PET textbooks and past PET papers was not very relevant for their students' futures and that 'topics focused on Japanese students would be better.'

A positive aspect of using the PET was expressed with this comment: 'PET has the 'Can do' list so that I can understand the aims clearly and provide lessons with clear ideas to the students.' Another teacher wrote, 'I also like the current exam system because it allows us to evaluate not how well our students understood the textbook(s), but how well their English proficiency has improved.' Overall, the questionnaire results indicated that teachers had mixed feelings towards the PET program, with a slight tendency towards dissatisfaction.

### Teacher interview results

Generally, all four interview participants admitted that their experiences of teaching toward the PET were mixed. It was noted that the PET provided a good framework to start from, especially for teachers who had never taught at a tertiary institution before. At many other universities in Japan, new teachers are given very few guidelines within which to conduct their classes, so this aspect was a perceived benefit. Generally, the speaking and listening sections were very well evaluated, while the writing and reading sections were not as well regarded.

*Test format*

Teachers seemed to rate the speaking test very well, at least in terms of the format. Mac (all names used are pseudonyms) said 'The oral test itself is very good.. It's nice not being one on one. Being one on two, I think, is better for the students.' (For a discussion of paired vs. singleton speaking tests, see Foot, 1999). Part 2 of the writing test was also evaluated well, being viewed as relevant to the real world. In this section, students write a short communicative message (35-45 words). They are told to whom they are writing and why, and must include three content points.

The general communicative format of the test was also well regarded. Atsuko noted:

> 'PET aims to develop communicative competency, right? And … students want to develop their communicative competence rather than translation

> skills or grammar knowledge. So the PET goals and student aims match. Right? So that is good. And the activities that I do in class are communicative, and PET coursebooks have lots of communicative activities. So that is also good. Students enjoy those communicative activities.'

*Test content*

Unlike the test format, the test content was not so highly evaluated. Regarding the speaking section, Mac found that it was difficult to create practice examples that were relevant and engaging. He said 'The biggest problem is coming up with good example situations for the area they're going to be tested on.' Relevance was a common area of concern for all teachers. Lucy said, 'I wonder about the relevance of the topics sometimes. You know, as far as, why are we teaching what we teach?' Yuka questioned whether students themselves understood the relevance: 'They can read the textbook and think about the questions and answer them. But I wonder if they see the point, like, why do they have to do it? What are they supposed to learn from that?'

Similar feelings were expressed regarding the extended writing section. One respondent stated:

> 'Unfortunately, there aren't situations when students need to write letters to people in English. Maybe even less so when they need to write a short story… So, I guess maybe the PET is designed for Europe, where English is a lot more out there… Japan is more remote, and the writing part is better suited to Europe.' Atsuko also mentioned that 'it would be nicer if [the test materials] are more localized to the Japanese context. It's better.'

In a similar vein, Lucy believed that students in Japan are disadvantaged when compared to students learning English as a second language in a European context. She believed that the PET content was more geared towards students in an ESL situation, and therefore not so suitable to the Japanese context. She commented:

> 'This is one of the huge differences between ESL and EFL. ESL students in an English environment overseas, studying with peers from different places can somewhat reasonably be expected to make a good guess based on their experience or their friend's experience/language learning base. EFL students don't have either advantage. Unless I spend hours and hours coming up with supplementary materials to help them be able to do this they don't really have a chance to figure it out.'

All participants saw the lack of English varieties in the PET as a problem. Mac said 'If they really expect the PET to be a world test, they should use world Englishes.' This perception remains despite Cambridge English exams aiming to cover all major varieties of English; they are 'designed to be fair to users of all nationalities and linguistic backgrounds' (Cambridge ESOL 2011c).

Another major concern was that the PET is a general English test, which means that students cannot know specifically what to study beforehand. Both Mac and Lucy used the word 'frustrating' for students. 'I think it's unfair on them', said Mac. 'And it's frustrating for them to not quite know.. what they're going to be tested on and how to prepare for the test.' Lucy said this was also a problem for teachers:

> 'As teachers, we really need to know exactly what's going to be on the test by the end of the semester. I think that would be a lot better than having a random sample of what it might be… As a teacher trying to prepare my students, no matter how hard I work, there's a very high possibility that there will be things on that test that we did not cover.'

However, despite this, Yuka saw the general nature of the test as empowering:

> 'For some students they may not know how to prepare for the exam. Because students often ask me, like, 'How should I study for the test?'. … But it's not really like that, right. If they want to get a good score on the test, they just have to study English. In this way, they have more choices, like how they can study English.'

The comments made by teachers during the interviews are individual perceptions based upon their experience of a specific PET program that forms part of their university's English language curriculum. No doubt, the specific program of study in which they worked influenced their opinions in no small way, and teachers may not have been aware of the exam support materials available at the time (see Lucy's comment above). As researchers we must therefore remain cautious about making any wider generalisations beyond the local context.

*Issues with the program*

It was quite difficult to separate teachers' attitudes towards the PET from their attitudes towards the university program of study within which they were teaching, as these were tightly woven together. Regarding the university program, one major issue was time. Every teacher interviewed felt that one 90-minute lesson a week was not enough time to prepare students adequately. Mac commented about Part 2 of the speaking test:

> 'There's a lot you can talk about with the picture. I mean, you could spend an entire semester on how to describe pictures and go over grammar forms. You could design a whole course around that if you wanted to. And so that is something that I see in our situation as a problem.'

There was some disagreement over the textbooks and other supplementary resources used in the program. Lucy noted, 'The textbook we're using right now I find very frustrating to use. I think the idea of the [PET] test is really good, and where it's aiming, but I don't really feel like the textbook gets us there.' On the other hand, Atsuko and Yuka appreciated the PET textbooks. Atsuko commented, 'Japanese published teaching materials are not based on CLT [Communicative Language Teaching]. They are more often based upon the grammar translation method, or the focus is on grammar. So the layout or color or design of PET materials is better.'

What participants did agree on, however, was that every teacher needs to make a special effort to adapt the PET textbooks and practice exam questions to be more engaging and motivating for the students. Yuka noted that, with the reading material, 'I have to kind of do something about it to make it more interesting or easy to understand.'

## Discussion

When considering the appropriateness of using the PET as a general proficiency test in a Japanese university, there was some difference between the attitudes of students compared with teachers. Generally, the PET was received more favourably by students. The background questionnaire indicated that students expected to study communicative English at university, and desired to improve their practical English ability. In this respect, the PET is an appropriate target for the needs of Japanese university students. After taking the PET, students said they enjoyed taking the test, they did not find it too easy or too difficult, and they believed it prepared them well for situations in their future life when they may need to use English.

Regarding the format and content of the test, students were generally positive. A large number believed that the PET accurately assessed their English proficiency, mainly because of the inclusion of a speaking section. Other more popular tests of general English, in use across Japan, do not assess speaking ability. English education in Japan generally is not focused on communicative ability, but rather on the ability to produce correct grammatical forms. Students recognise and appreciate the communicative nature of the PET. For these reasons, taking the PET exam had a motivating effect on the participants. After finishing the first PET exam, every single student indicated that they were motivated to continue studying. Although the perceived European flavour of the test caused consternation for a number of teachers, the students seemed to have no such misgivings and they felt the content was suitable for Japanese students.

While students seemed quite satisfied with the PET, teachers were a little more negatively disposed to the content of the test. The main criticisms centred around the perceived lack of world Englishes and the Eurocentric content. While students judged the test to be relevant for their future, teachers said that it was not; they desired more local (Asian) content. These findings are consistent with what Harwood (2007) found regarding the KET. On the other hand, the communicative nature of the PET was well regarded by the teachers. The speaking section was evaluated highly, as was the listening section.

Most of the negative comments from teachers related to the textbooks used which are not general English course books, but rather PET preparation textbooks. As previous research has shown, teachers do not like preparing students to take a test. When a semester-final test becomes the focus of every lesson, teachers feel like they are teaching 'test-taking skills' rather than 'English communication skills'. This issue was compounded by the fact that items which arose on the final test may not have been covered in class. Teachers said this was 'frustrating for the students', although the students themselves expressed no such frustration. The implications for any university considering developing such a program, is that perhaps teachers would be better served by using general English course books aligned to the CEFR, as opposed to a test preparation textbook, and that the nature of the PET as a 'general English' exam needs to be emphasised as much as possible. It needs to be stressed to the teachers that they are not 'teaching to a test', but rather, they are 'teaching general English'. The semester-final exam is merely the instrument used to gauge how much the students' level of general English has improved.

When examined closely, it appears that teachers are not so critical of the PET itself, as they are critical of a program of study based around a standardised test preparation textbook and focused towards a semester-final exam. One can imagine getting the same responses from teachers using a TOEFL preparation textbook with TOEFL as the final exam. Although space does not allow a full discussion here, readers should bear in mind the implications of using proficiency tests such as PET as achievement tests, each type of testing having differing purposes and outcomes (see Davies 1999:154). Many universities in Japan do not have a standardised program and unified curriculum; teachers are given a free hand to choose their own materials and make their own tests. It seems that there is some resentment at having to follow a unified course and this is likely to have informed the negative responses revealed in this study. There are also higher-level curriculum planning and policy decisions that informed the teachers' experiences and hence the results of this study. Language teachers world-wide face different demands depending on whether they are expected to follow a narrow curriculum to prepare for a specific test or whether they are allowed more flexibility which places the onus on teachers to provide opportunities for students to become familiar with a test within a general language learning program.

## Conclusion

With the growing influence of the CEFR, as well as the current popularity of communicative language teaching, the use of a CEFR-based general English test as a university semester-final exam is one option for curriculum planners. In the case of using PET in a Japanese university setting, this is indeed a viable option. Japanese students regard the PET favourably and they find it motivating and relevant. The greatest challenge to be overcome is the attitude of teachers, which tends to be more negative than positive. In order to empower teachers and give them an increased sense of ownership in the program, course directors would be well advised to choose a general English course book rather than a test preparation textbook. Perhaps even more important than this, however, is to foster an environment of peer support and collaboration amongst teachers. When the teaching staff do not share in the goals and vision of the program, small annoyances become major hurdles. The perceived Eurocentric nature of the PET and the subsequent lack of Asian content are serious shortcomings in the eyes of many teachers, although students expressed no such concerns. These perceived shortcomings could be overcome by teacher training or the creative use of supplementary materials in the classroom, but this requires an investment of time and effort that, unfortunately, some language teachers are unwilling or unable to make.

Concerning the PET exam itself, this research suggests that teachers preparing students for the PET in Japan would appreciate more varieties of World Englishes in the exam and both teachers and students would welcome topics of relevance to Asian students. However, it should be acknowledged that this desire is fraught with difficulties. The PET is an exam for international candidature, and as such is based on standard forms of English (US, Australian, etc.) in order to avoid giving advantage to a particular language or cultural group.

The results of this study indicate that the PET, in itself, is an appropriate assessment tool to be used in the Japanese university setting. If the CEFR continues to strengthen its influence around the world and especially in Asia, program directors will be faced with the goal of how to align their curricula with this framework. Further studies of the PET, and indeed the KET and FCE, in tertiary contexts would be crucial in helping us move toward this goal.

## References and further reading

Alderson, C and Hamp-Lyons, L (1996) TOEFL preparation courses: a study of washback, *Language Testing* 13 (3), 280–297.

Buck, G (1988) Testing listening comprehension in Japanese university entrance examinations, *JALT Journal* 10,15–42.

Cambridge ESOL (2009) *Cambridge English: Preliminary: Handbook for Teachers*, retrieved 16 January 2012 from: www.teachers.cambridgeesol.org/ts/digitalAssets/114933_pet_handbook.pdf

Cambridge ESOL (2011a) *Statistics for individual exams: Preliminary English Test (PET)*, retrieved 25th October 2011 from: www.cambridgeesol.org/what-we-do/research/grade-stats/2009/pet.html

Cambridge ESOL (2011b) *Studying and working abroad*, retrieved 25th October 2011 from www.cambridgeesol.org/sector/study-work-abroad/index.html

Cambridge ESOL (2011c) *Cambridge English Preliminary*, retrieved 24th October 2011 from: http://www.cambridgeesol.org/exams/general-english/pet.html

Capel, A and Nixon, R (2003) *PET Masterclass*, Oxford: Oxford University Press.

Capel, A and Sharp, W (2005) *Objective KET*, Cambridge: Cambridge University Press.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment*, Cambridge: Cambridge University Press.

Davies, A (1999) *Dictionary of Language Testing*, Studies in Langauge Testing volume 7, Cambridge: UCLES/Cambridge University Press.

Foot, M (1999) Relaxing in pairs, *ELT Journal* 53 (1),36–41.

Hamp-Lyons, L (1998) Ethical Test Preparation Practice: The case of the TOEFL, *TESOL Quarterly* 32 (2), 329-337.

Harwood, C (2007) *Washback and the Cambridge ESOL Key English Test Speaking Component: A study from Japan*, unpublished Masters dissertation, University of Leicester, UK. Retrieved 15th March 2011 from: http://nus.academia.edu/chrisharwood/Papers/280354/Washback_and_the_Cambridge_ESOL_Key_English_Test_Speaking_Component_a_study_from_Japan

Lewthwaite, M (2007) Teacher and Student Attitudes to IELTS Writing Tasks: Positive or Negative Washback? *Learning and Teaching in Higher Education: Gulf Perspectives* 4 (2). Retrieved 15th March 2011 from: http://www.zu.ac.ae/lthe/lthe04_02_03_lewthwaite.htm

McCarthy, M and O'Dell, F (2005) *English Collocations in Use*, Cambridge: Cambridge University Press.

Naylor, H and Hagger, S (2004) *Insight into PET*, Cambridge: Cambridge University Press.

Raimes, A (1990) The TOEFL Test of Written English: Causes for concern, *TESOL Quarterly* 24, 427–442.

Read, J and Hayes, B (2003) The impact of IELTS on preparation for academic study in New Zealand, *British Council/IDP Australia IELTS Research Reports* volume 4, 153–206. Available online from http://www.ielts.org/researchers/research.aspx

Redman, S (2003) *English Vocabulary in Use: Pre-intermediate and Intermediate*, Cambridge: Cambridge University Press.

Shizuka, T (2002) *Eigo tesuto sakusei no tatsujin manual* [A manual to become a master of making English tests], Tokyo: Taishukan Shoten.

Shohamy, E (1992) Beyond proficiency testing: A diagnostic feedback model for assessing foreign language learning, *Modern Language Journal* 76, 513–521.

Shohamy, E, Donitsa-Schmidt, S and Ferman, I (1996) Test impact revisited: washback effect over time, *Language Testing* 13 (3), 298–317.

Spratt, M (2005) Washback and the classroom: The implications for teaching and learning of studies of washback from exams, *Language Teaching Research* 9, 5–29.

Torikai, K (2010) *Eigo koyo-go wa naniga mondaika* [What are the problems with making English an official language (in Japan)?], Tokyo: Kadokawa Shoten, 47-76.

Wakabayashi, S and Negishi, M (1993) *Musekinin na tesuto ga ochikobore wo tsukuru* [Badly-made tests make left-behind students], Tokyo: Taishukan Shoten.

Watanabe, Y (2000) Washback effects of the English section of Japanese entrance examinations on instruction in pre- college level EFL, *Language Testing Update* 27, 42–47.

# FCE exam preparation discourses: insights from an ethnographic study

**DINA TSAGARI** DEPARTMENT OF ENGLISH STUDIES, UNIVERSITY OF CYPRUS, CYPRUS

## Introduction

So far various research studies have looked at test washback, that is the influence of tests on teaching and learning (Alderson and Wall 1993:214). Studies of test washback have investigated the effects of local, national and international standardised language tests in various educational contexts focusing on a variety of 'participants' and 'products' (Bailey 1996), e.g. teachers and teaching, learners and learning, teaching materials, attitudes towards testing, etc. (see Tsagari 2009 for detailed review of the literature). The studies resulted in varying conclusions about the absence and presence of washback and its degree (positive or negative) mainly due to different learning contexts, teachers' beliefs, research methods used and stakes of the tests under study.

With regard to teaching methodology, which is of concern to the present research, the studies have shown that washback on how teachers teach is still unclear and complex. The studies follow a cline indicating presence of washback (Munoz and Alvarez 2010, Saif 2006, Shohamy 1993, Shohamy, Donitsa-Schmidt and Ferman 1996, Stecher, Chun, and Barron 2004) to absence of washback (Qi 2005, Wall 2005, Wall and Alderson 1993, Wesdorp 1983). It is also interesting to note that the studies that found evidence of washback on teaching also found large differences in the way teachers teach towards the same exam (Alderson and Hamp-Lyons 1996, Burrows 2004, Cheng 2005, Hayes and Read 2004, Qi 2005, Watanabe 1997). In addition to the above, studies of test washback have been based on data such as pre- and post-test performance scores, interviews and questionnaires with candidates and teachers (Becker 1990, Elder and O' Loughlin 2003, Green 2003, Hayes and Read 2004, Rao, McPherson, Chand and Khan 2003, Robb and Ercanbrack 1999). The scope of such studies has been restricted in that they considered aspects of preparation programmes on particular skills or studied the effects of specific test preparation programmes on scores and the influence of tests on teacher perceptions or attitudes, rather than examine details of teachers' instructional behaviours and provide descriptions of classroom practices. However, the studies stressed the need to be clear about which features of classroom behaviour to observe when researching test washback on teaching as this is not an inevitable or universal phenomenon. Future research, therefore, needs to look at the influence of tests on teachers' methods more closely by employing descriptive studies of classroom instruction of candidates preparing for high-stakes tests to allow researchers to investigate some of the apparent contradictions in the findings to date (Spratt 2005). Green (2006) also recommends the use of 'more sensitive instruments' such as in-depth interviews and classroom observations when researching test washback (stressed also in Mickan and Motteram 2008).

## The research study

This article reports an empirical study into the instructional practices used by English language teachers who prepare candidates for Cambridge English: First (*FCE*)[1] administered by Cambridge ESOL in the context of Cypriot private language schools, known as 'frontistiria'.More specifically, the study investigated the features that characterise instruction in the *FCE* preparation classes to better understand the conditions for development of language skills tested in the *FCE* and explored implications for exam preparation programmes. This is of importance in language testing, as the study of the relationship between instruction and language skills measurement has the potential to contribute to the external credibility of a test (Brindley and Ross 2001).

The study addressed the following research questions:

1. What is the nature of the instruction in *FCE* preparation programmes?
2. How do the instructional practices observed in *FCE* preparation programmes relate to the test requirements?
3. What are the implications for the preparation for the *FCE* examination?

In line with the view that the teaching and evaluation of language activity needs to be done at a discourse level (Mickan 2000, Van Lier 1988) and the number of research studies that have emphasised classroom discourse experiences of candidates as important influences on test performance (Mickan 2003, Mickan and Motteram 2008, Mickan and Slater 2003, Mickan, Slater and Gibson 2000, Munoz and Alvarez 2010), the present investigation documented and analysed teacher and student discourses in *FCE* exam preparation courses as valuable resources of information of instructional discourse in the preparation of *FCE* students. The study also drew on the work of Mickan and Motteram (2008) and employed a qualitative orientation to the analysis of data which was expected to provide insights into learners' experiences of instruction (Unsworth 2000) and testing (Mickan et al 2000, Torrance 1995, Weeden, Winter and Broadfoot 2002).

### Participants

Given the constraints of the local educational context[2], a decision was taken to collect data through interviews with *FCE* teachers and classroom observations from more than one frontistiria. In the course of the study, three frontistiria owners agreed to allow teachers' interviews and observations; therefore four teachers and 26 students preparing for the *FCE* exam participated in the study. Fifteen lessons were observed and tape-recorded, which produced 24 hours of observations (see Table 1).

**Table 1: Overview of classroom observations**

| Schools | Teachers observed | Lessons observed | Hours of observation |
|---|---|---|---|
| School A | T1 | 3 | 6 |
| | T2 | 3 | 4.5 |
| School B | T3 | 6 | 9 |
| School C | T4 | 3 | 4.5 |
| **Total** | **4** | **15** | **24** |

The data collection period (teachers' interviews and observations) took place early in the preparation cycle, e.g. between end of September till mid of November 2010 (students were all preparing to take the *FCE* exam in June 2011). Consequently, any exam influence observed would serve to underscore the influence of the *FCE* on language learning in these classes (Bailey 1999, Messick 1996).

### Data collection and analysis

The four teachers were interviewed on a number of aspects of exam preparation as these were discussed in the literature

[1] http://www.cambridgeesol.org/exams/

[2] Preliminary meetings with frontistiria owners, directors of studies and teachers conducted May–July 2010 in the wider area of Nicosia, Cyprus, revealed that it was difficult to obtain permission to observe teachers for an extensive period of time. Owners and teachers explained that students and teachers would feel uncomfortable being observed for a long period of time. They also stressed that the presence of an outside observer in their *FCE* classes would have a negative effect on the preparation process and that parents would not give their consent. However, the institutions would be content if a smaller number of visits and observations were made.

and laid out in the *FCE Handbook for Teachers* (UCLES 2007). For this purpose an interview worksheet was designed which covered the following areas:

**Table 2: Overview of FCE interview worksheet**

| Part | Title | Number of questions | Contents |
|---|---|---|---|
| Preliminary info | | 3 | School/teacher information |
| I | Teacher biodata | 4 | Teaching experience, qualifications, FCE training, years of FCE preparation |

Classroom observations comprised the core data of the study. Other than tape-recording the lessons observed, a specially designed observation schedule was used that focused observation on aspects of the lesson that related to teachers' practices towards the *FCE* preparation, e.g. it described each classroom activity in rich detail and allowed the observer to state whether the activity was related to *FCE* tasks. The observation schedule also captured specific references by the teacher or students to *FCE* content, format, and other exam requirements.

This data along with copies of the learning materials used in the *FCE* classes observed, was used as supplementary data to illustrate teachers' instructional practices.

The recordings collected from classroom observations were transcribed verbatim to provide an accurate record of the classroom discourse produced in the lessons recorded. For the analysis and interpretation of the observational data a special statistical package, e.g. ATLAS.ti 5.0 (Muhr 2004), was used. A Coding Scheme[3] was used to analyse the observational data that was based on the research questions, the *FCE* handbook (UCLES 2007) and findings from the research literature reviewed. The Coding Scheme applied to the transcripts comprised five parts (as many as the papers of the exam) and several categories, e.g. materials, tasks, topics, exam advice, etc.

Several validation checks took place during the process of data collection and analysis. For instance, transcription and analysis of the first lesson recording gave the opportunity to try out the Coding Scheme on real data and refine it for subsequent data collection. In addition, to ensure internal consistency of the analysis of the observations, the researcher, following Cohen, Marrion (2000) suggestion, used peer examination of the data. This aimed to serve as a reliability check of the analysis of the present researcher. The transcript of the lesson and the Coding Scheme were given to a colleague to use who had extensive experience in teaching exam-oriented classes. This strengthened the validity of the interpretation of the data as the results indicated that both analyses of the data were similar.

### Interpretation of the results

For the interpretation of the data, the study employed a sociocultural theory perspective. This has had a significant impact on the interpretation of classroom experiences and practices and on the analysis of the development of language skills (Kramsch 2002, Lantolf 2000, Lantolf and Thorne 2006, Mickan 2006a, 2006b). This perspective was expected to add a new dimension to the analysis and interpretation of observational data in empirical washback studies. Overall, sociocultural theory suggests that human behaviour is a result of the integration of socially and culturally constructed forms of mediation into human activity (Lantolf 2000). Swain, Kinnear and Steinman (2010:x) stress that, according to Vygotsky, 'the source of learning and development is found in social interaction rather than only in the mind of the individual'. In line with this way of thinking, teacher discourse would be related to the sociocultural reality of the context under study. Therefore, the findings from the classroom observations were interpreted and reflected upon through the realities of the local society and culture these occurred in. Factors influencing teachers' activity such as the place of English as a foreign language in Cyprus, the importance of *FCE* in the Cypriot society and culture and the role of private institutes in the educational EFL context in Cyprus were taken into consideration when the data was analysed.

## Presentation of results

The following sections present the analysis of the data with direct extracts from the transcripts to illustrate the points being made. At the end of each extract there is a code which represents the teacher and the lesson observed, e.g. T1 (the first teacher), L1 (the first lesson conducted by the teacher). Teacher and student names are anonymised. When extracts are italicised, these are translations of teachers' discourse from Greek into English made by the present researcher. The remaining extracts represent the teachers' exact words in English.

### Teacher profiles

*Teacher 1*

The interviews and classroom observations showed that Teacher 1 established a good rapport with her students by making them feel part of the learning process, e.g.

> Teacher 1: So before we finish our lesson today I would like to thank you for your co-operation.
> (T1, L2)

English was used in the classes observed by Teacher 1 and her students. The teacher constantly provided her students with information and advice about the *FCE* exam and offered students various opportunities for *FCE* practice. In the classes observed, Teacher 1 assumed the role of the facilitator, allowing students to find their ways to success (Brown 2004). She provided guidance but promoted student independence at the same time by asking them to assess themselves and their peers. As the teacher explained during her interview, she hoped her students would 'acquire self-awareness of mistakes' and 'feel trustworthy'. The following extract shows the teacher's effort to raise her students' awareness of language errors:

[3] This can be made available on request.

Teacher 1: Good effort! OK. I want you to go back to your checklists. It is very important. It is time to assess yourself . . . put a score out of ten . . .
Teacher 1: Why are we doing this, Student 1?
Student 1: To become critical.
Teacher 1: Why we need that?
Student 2: Become self-aware from our mistakes . . .
Teacher 1: I trust you enough to become critical.
Student 1: I can't.
Teacher 1: I trust you . . . I can help you . . . If you have a problem, you should let me know.

(T1, L3)

Teacher 1 was flexible in her teaching. She did not hesitate to change her lesson plans when she felt it would benefit her students. As seen in the following extract, the teacher first informed students about her decision to postpone the presentation of language items (phrasal verbs) and explained that the reason behind her decision was to facilitate their learning:

Teacher 1: No! Look, look. Phrasal verbs are the important part. You had to spend at least one week at each part of phrasal verbs. That's why I postponed this phrasal verb part . . . to have a longer time between the previous phrasal verbs to these ones, so that you can have enough time to study and learn the previous ones. OK? And not following the lesson plan because I want to give you more time to study. I don't want extra time to be wasted.

(T1, L1)

*Teacher 2*

Teacher 2 offered advice on how to prepare for various tasks of the *FCE* exam without making specific reference to the exam itself. There was lack of rapport between the teacher and the students who remained silent in class most of the time. What characterised the classes of the particular teacher was the infrequent use of L2. Unlike Teacher 1, Teacher 2 used her mother tongue frequently, and so did her students, to provide explanations or advice to students:

Teacher 2: *You have to be more careful with your exercises and think of what the particular tense shows. You still haven't understood the reasons why we use each tense. OK? . . . OK, who will explain the meaning of this in Greek?*

(T2, L2)

The teacher was 'the manager of the class' (Brown 2004) with lessons that were delivered according to plan and tasks that were chosen in advance. Teacher 2 directed activities and students to the goal of her lessons:

Teacher 2: OK, let's start by correcting our homework . . . Let's do exercise A, complete the sentences using past simple or past continuous active or passive, this one, OK? . . . OK, first I want you to tell me all the keywords and expressions you found, highlighted or underlined.

(T2, L2)

*Teacher 3*

Teacher 3 provided tips and advice about the FCE exam by referring directly to the exam. She would resort to L1 from time to time to provide explanations of language use but her students did so more often.

Teacher 3: *. . . deeply is an adverb, -ly deeply. Deep is the adjective . . . We add –ly and it becomes deeply.*

(T3, L1)

Teacher 3 made jokes during her lessons. This was appreciated by her students and created a pleasant atmosphere in the classroom. However, the teacher also made some demeaning comments, too. Furthermore, just like Teacher 2, Teacher 3 was the 'manager' of the class. She provided students with opportunities to practise for the exam and become familiar with its format:

Teacher 3: The first exercise. . . the first listening exercise on page twenty-four. . . focuses on . . . the exercise you will have on the exam . . . it's the part four exercise, OK? Part four is where you will hear an interview or a conversation between two people, OK? And you will have to answer questions which concern the interview, OK? So the questions concern the whole interview it's not just . . . you won't listen to eight different conversations like the previous one, OK?

(T3, L3)

*Teacher 4*

Teacher 4 mainly focused on delivering her lessons and avoided any kind of remarks or jokes during her lessons. She used L2 to talk to students, give advice about the exam and explain language structures and use:

Teacher 4: Yes but we have 'seen'! Is it a verb? We need a verb . . . 'have seen'. What is 'seen'? It's a past participle. Alright? What tenses do we use in past participle?. . .

(T4, L2)

Teacher 4 tried very hard to support her students and explain what they had to do in the tasks assigned, in the following example by explaining the exercise individually to each student and giving them time to do it.

Teacher 4: Do it like this . . . it's the same. OK? . . . This is a phrase with 'with', OK? Is that a phrase? Using 'with'? OK? It means do something and not . . . same thing. Exactly the same thing . . .

(T4, L3)

### Preparing for the FCE papers

The sections that follow demonstrate how each skill was handled by teachers in terms of test practice and advice.

*Reading comprehension*

Overall, when teachers worked on reading in class, they had their students do reading tasks such as selected response, which is the type of task used in the Reading paper of the exam. No other types of reading tasks were used (e.g. open-ended response) other than the *FCE* reading tasks. In addition, each teacher offered advice, albeit somewhat differently, to her students on various aspects of the Reading paper. For instance, Teacher 1 was concerned about her students' approach to reading. In the following extract, the teacher, using an inductive approach (asking questions), reminds them of what needs to be done when working on reading tasks:

Teacher 1: Easy? Good. Do you have anything to ask me? In terms of the procedure? Reading the text? Underline the key points? Go through it again? Did you guys do that? Come on! Be honest! [calls on a student] What did you do? In terms of understanding it?

(T1, L2)

Teacher 1 prompted her students to read outside class time as well in an effort perhaps to make them independent and confident readers:

Teacher 1: The magazines are yours. Make sure you write your name inside. OK? It's an extra source you can read when you go home.

(T1, L3)

Teacher 2 did not do a lot of reading practice nor did she offer advice to her students when working on reading. Instead she did work on vocabulary and instructed them to look up new words in their companion books[4]:

Teacher 2: Ok before you read the questions, I want you to open your companion book and look for new vocabulary there please. OK? Elena?

(T2, L4)

Contrary to this, Teacher 3 advised her students not to pay attention to unknown words:

Teacher 3: If you have unknown words try not to be affected please . . . in the exam you should try to guess the meaning from the text or do not pay attention to them at all.

(T3, L1)

Teacher 3 also recommended three approaches when reading for the exam and advised students to choose the one they felt was more suitable for them. Students were asked to read the questions first, underline keywords and then read the text and go back to the questions in search of the answers, e.g.:

Teacher 3: I've told you a lot of ways how to . . . read these kinds of texts for the exam and how to find answers. OK? Now I want you to choose which way is more convenient to you. First way is . . . OK . . . we have read the questions, we know the keywords, we know the words which will help us focus on the answers so now you can go back, read texts A, B, C, and D and . . . then go back to the questions and try to find the answers . . .

(T3, L1)

According to the second approach, students were asked to read the questions after each text and decide whether these related to the text. The third approach recommended was to read each text to get the main idea first. Once students did that, they were asked to go back and answer the questions.

Finally, Teacher 4 informed her students about the format of the Reading paper and advised them how to prepare for it. In the following example, she explains how to work on Part 3 (e.g. multiple matching) of the Reading paper:

Teacher 4: So, OK? Look at the board. Here I have some options. A to H. I want you to match them with the paragraphs in your text. So . . . each title goes to a paragraph. Alright?
Student 3: *Miss, there are more titles than paragraphs!*
Teacher 4: Always! OK? There is always one extra. OK?

(T4, L3)

*Writing*

To prepare their students for the Writing paper, teachers' practice covered out-of-class and in-class writing as well as analysis of model compositions and in-class correction of student compositions.

Students were also offered advice on test-taking techniques for writing. For example, Teacher 1 reminded students of the writing genres required in the exam:

Teacher 1: OK, so . . . what is our genre? What are the kind of texts you are going to write about in the exam?
Student 4: Romantic.
Teacher 1: In general. . . What about this one? It's what? A report or?
Student 4: No. It is a story.

(T1, L3)

Teacher 1 also corrected the writing tasks in class and encouraged students to be critical of their own and their fellow students' writing. Actually, every student was called to read their story and offer their peers suggestions for improving their written compositions:

[4] Companion books, part and parcel of the *FCE* textbook packages, are glossaries of unknown vocabulary found in the Students' book accompanied by explanations illustrated with examples and, occasionally, by practice tasks.

Teacher 1: Let's get some feedback. Student 5, you need a pencil? Let's go . . . Speak to her.
Student 5: Me?
Teacher 1: Yes, . . . go to the features, be specific, try to think harder, think about her story. Is there something specific that you like?
Student 5: It has clear meaning. And I like it.
Teacher 1: Try to think harder Student 5.
Student 5: Actually, I like the story when it ends bad.
Teacher 1: Yes you do?

(T1, L3)

Teachers 2 and 4 stressed the usefulness of specific grammatical features such as the use of past tenses in writing as in the example below:

Teacher 2: *OK they are past tenses. They are very nice tenses to use when we write a story, when we narrate an event. And that is what we are going to do today, we'll see how we write stories, we'll look at the plan we follow, the words we use.*

(T2, L3)

Teachers 2, 3 and 4 presented students with model compositions which they analysed together with their students to familiarise them with the requirements of the Writing paper, e.g.:

Teacher 2: Look at the model please and tell me . . . which three past tenses are used? Look at the model composition and find examples from past tenses.

(T2, L1)

The same three teachers also encouraged students to follow the format of model compositions. The teachers drew students' attention to the structure of the compositions and gave advice about composition planning. In the following extract, Teacher 2 explained the importance of the development of a clear and coherent storyline following the exam specifications:

Teacher 2: . . . Four paragraphs. And you must do exactly the same thing as in the exam model. Opening paragraph, paragraph 2 . . . making the end leading towards the climax, paragraph 3 . . . you must say what the climax is, paragraph 4 . . . try to say how the story ends . . . *how you felt.*

(T2, L3)

Teacher 3 also helped her students to organise their ideas. In the following extract, students working on the model composition of an email were advised to practise writing appropriate opening and concluding paragraphs and then use the guidelines provided in their textbook. This was also recommended in the *FCE* handbook (UCLES 2007:21):

Teacher 3: Right, let's read the model email below and see what the friends' suggestions are . . . Organise your ideas, greeting, 'Dear Mr. Stone', opening paragraph, give your reason for writing. What's the reason here? . . . The main body, two or three paragraphs relate to the point in the email and notes . . . So look at the second paragraph . . . Closing paragraph four, end with appropriate closing. How does Laura close her paragraph?

(T3, L6)

Teacher 4, through scaffolding questions, tried to help students become aware of the appropriate style and tone of the writing genres (e.g. magazine article) needed in the exam (UCLES 2007:22):

Teacher 4: So, why should we write in a school magazine? Why? Does it play a role or something? Does it mean anything? Why does it say 'write a story for a school magazine' and not for a scientific magazine or something else? Yes?
Student 6: Because we write it more formal.
Teacher 4: Formal? What about here? Is it a formal story? No! It's informal OK? We need to have an informal style . . . OK! So, what sentence must you include in your story?

(T4, L2)

All four teachers offered advice on test-taking techniques for writing. For example, Teacher 1 reminded students of task length:

Teacher 1: OK . . . length . . . what was the word limit? 200 words? OK? . . . Nicolas?

(T1, L3)

Teachers 2 and 4 had their students brainstorm in order to help them cope with writing. This practice is in line with the *FCE* handbook guidelines (UCLES 2007:77), e.g.:

Teacher 2: *When we have a topic we always try to think of various ideas* . . . OK, so brainstorm ideas. This is your planning phase. Brainstorm ideas. You might want to think about where and when the holiday was, who went, what happened and why it was the best holiday ever. If you answer these questions you will finish your compositions, OK? You must answer all these questions in your story. Now, organize these ideas. The plan is in exercise 4. How many paragraphs?

(T2, L3)

Teacher 2 stressed the importance of using correct tenses and time expressions so that the flow of ideas would be logical and easy to follow, which is also in accordance with the general preparation advice offered in the *FCE* handbook (UCLES 2007:20):

Teacher 2: So when you write your own composition I would like to see past continuous, past simple and past perfect in your stories. Another thing that I would like to see is time expressions. Because I want to know what happened first, second, after . . . after that, later, in the end OK? So, I want to know the sequence of events. OK, look at the time expressions on your handout and tick the ones that are used in the model please.

(T2, L3)

Teachers 3 and 4 also emphasised the importance of task instructions and urged their students to read the rubrics preceding the writing tasks carefully as recommended in the *FCE* handbook (UCLES 2007:21):

Teacher 3: . . . you lose points because here you have to answer these questions . . . there are questions and you should answer . . . you should include everything, every single note of the instructions in your composition.

(T3, L6)

Teacher 4 had her students mark the compositions of their peers but did not use the marking criteria for writing as specified in the *FCE* handbook (UCLES 2007:28). She did so based on an impressionistic evaluation of the quality of peer texts:

Teacher 4: OK, yes. Good. So. What do you think? If you were teachers? If you give a mark to the story what will it be?
Student 7: 20 out of 20.
Teacher 4: You have to say why. Don't forget. Yes?
Student 7: 9 of 10 because it is a little fantastic . . .
Teacher 4: Imaginative?
Student 7: Yes
Teacher 4: OK. Alright, Student 8, what do you think?

(T4, L2)

*Use of English*

The four teachers placed a lot of emphasis on the skills needed for the Use of English paper. Some of them did so by referring directly to the *FCE* exam while others did not. In the observational data, numerous extracts provide evidence of the teachers' efforts to raise awareness of test features. In these extracts, the majority of the teachers are depicted as 'knowledgeable' instructors giving advice and explaining test-taking techniques. For example, in the following extract Teacher 1 gives advice on how to handle cloze tasks (Part 2, Use of English paper). The teacher advises students to think of specific categories of words missing when working on such tasks:

Teacher 1: But let me tell you something. That will be very handy in the exam. The thing is that when you have a cloze test where you have to guess the words, usually the words are not difficult words. I mean they are not really words of vocabulary. They are linking words . . . let's say . . . modal verbs or auxiliaries. I mean prepositions or collocations here are also very important . . . So don't go for very special words. OK? Don't look for very special words. It's a preposition, an auxiliary or a collocation.

(T1, L1)

However, Teacher 4 followed a different approach. In the following extract, Teacher 4, through scaffolding, tried to help students think of the categories of the words missing in a cloze task:

Teacher 4: OK. Now, you must look at which words are missing. Is it a verb? Is it an adjective? Is it an adverb? Is a preposition? What?

(T4, L1)

Teachers 3 and 4 proposed reading cloze passages once before they filled in the words missing so that they can gain an understanding of the overall meaning of the texts:

Teacher 3: So we always, always, always read the open cloze text and then try to find the answers. OK? You have to understand the meaning of the text . . . the general meaning . . . all right?

(T3, L2)

Teacher 3 worked on Part 3 ('word formation') in her classes. She had students produce various forms of specific words orally as in the following exchange:

Teacher 3: Noun?
Student 9: Development
Teacher 3: mmm . . . Adjective?
Student 9: Developing?
Teacher 3: And?
Student 9: Developed?

(T3, L1)

During the observations, Teacher 2 did not ask students to practise Use of English tasks in class. Instead, she asked students to work on the tasks at home and then corrected them in class. She also advised students to look up unknown words in the dictionary:

Teacher 2: Didn't you look it in the dictionary? I told you. When you do the Use of English at home and you find an unknown word what is your reaction? You have to look them up in the dictionary.

(T2, L3)

The teachers, overall, reassured their students that this part of the exam was not difficult. They wanted to minimise students' stress and raise their confidence before the exam. Teachers also provided students with set phrases to help them face the requirements of this paper, e.g.:

Teacher 3: *. . . but when you see the task you will see it is not as difficult as you think it is . . . I am sure you will do very well in this . . . So . . .* here are some phrases which might help you all the way through the part three.

(T3, L4)

Phrasal verbs, in particular, were considered important for this part of the exam. In the following extract, Teacher 1 stressed their importance:

Teacher 1: Look, look! Phrasal verbs are the important part in the Use of English. You had to spend at least one week at each part of phrasal verbs.

(T1, L1)

*Listening*

During the classroom observations, Teachers 2 and 3 worked on listening tasks for the exam and offered various types of advice to their students. However, no listening practice was observed in the classes of Teacher 1 and Teacher 4. Teachers 2 and 3 devoted a lot of time to raising students' awareness of Listening test features. The teachers explained the requirements of specific listening tasks and made general comments about the Listening paper. In the extract below Teacher 2 provided information about the format of the Listening paper by discussing the role of distractors in Parts 1 and 4 (multiple-choice tasks):

Teacher 2: *. . . this type of questions are called distractors. They are specifically designed to distract your attention to check if you can understand what the text is about and choose the correct answer, OK?*

(T2, L2)

She also informed her students about what to expect and advised them not to be surprised if the questions they read in their exam papers did not match the language used in the audio recording:

Teacher 2: *. . . the questions are not identical to what you will hear. What you hear will be paraphrased.*

(T2, L2)

As in the case of reading, Teacher 3 offered two ways of approaching Part 4 of the Listening paper. According to the first one, students were advised to guess which answer might be the right one before they listened to the recording. Then they were asked to listen and check if they guessed correctly. According to the second approach, students were advised to read the multiple-choice options and choose their answers after listening to the recording. Teacher 3 also prompted her students to choose the one they thought was more suitable for them and emphasised the need for time management when working on the listening tasks:

Teacher 3: You will just have forty seconds to a minute to read the questions . . .

(T3, L3)

In a later lesson, Teacher 3 informed her class that each listening part should be handled differently and that varying strategies apply. She stressed the difference between Part 4 and Part 1 as follows:

Teacher 3: Part 4 was an interview and all the questions were about the interview. This part, part 1, talks about eight different situations, . . . for questions 1–8 choose the best answer a, b or c. Now, we cannot predict the answer here like we did in part 4. We can't predict the answer because it could be anything. OK? . . . So we're just going to read the questions, read the choices and go along as we listen to it, OK? You have to justify your answers in this part so be careful.

(T3, L4)

Attention was also drawn to the importance of words occurring before and after the listening gaps in Part 2:

Teacher 3: So we are looking for an adjective, OK? So, we will have in mind that we are looking for an adjective for number nine. Now, as soon as we listen to an answer and we write the answer down, we immediately go to the beginning of the second sentence and we wait, OK?. . . So we can listen to the beginning of the second sentence of the key words we will underline.

(T3, L4)

*Speaking*

Speaking practice was observed in the classes of Teacher 1 without being directly linked to the requirements of the exam. On the other hand, speaking practice directly targeted the FCE in the classes of Teachers 2, 3 and 4. It is also worth noting that teachers mainly used L2 when working on the speaking parts of the exam.

Teachers 2, 3 and 4 mainly offered advice on using appropriate language. Models of speaking were used by these three teachers. The models used aimed to set an example on how students were expected to perform in the speaking part of the exam, for example:

Teacher 3: Listen to two students doing the first part of the test. I want you to listen . . . listen to their conversation . . . and . . .. they will disagree on four items. Which items?

(T3, L5)

Teachers explained various task features related to the Speaking paper. In the example that follows Teacher 2 explained how students should approach Part 2 with regard to the stimulus (pictures) that will be provided by the interlocutor:

Teacher 2: OK, let's do some speaking. OK, next page, page 23, speaking, what do we have to do here? You have some pictures in front of you, you have to compare two pictures. OK, you have to answer some questions about the pictures . . . Where is the question? Can you see it? The question is written above the pictures, OK? How long have you got to talk about the two photographs? Do you know? You have to speak for one minute, OK? Here are the photographs . . . describe the photographs . . . and you have to speak for one minute.

(T2, L2)

Classroom activities which involve students working in pairs took place in the classes observed following the guidance offered in the *FCE* handbook (2007:78). For example, Teachers 2 and 3 had students practise in pairs for Part 2 (picture comparison):

Teacher 3: OK, now I would like you to work with your partners and compare the photographs, use the vocabulary that will help you. One will be student A and the other B. Decide which one are you going to talk about: this or this? . . . OK stop time is up. I have to remind you what I have said about the photographs. How long are you allowed to speak and compare the photographs?

Student 10: One minute.

(T3, L2)

Overall, teachers followed the suggestions of the *FCE* handbook (2007:20,55) with regard to time management. They informed their students of the amount of time allocated to each task:

Teacher 3: So . . . I was saying that you should divide your time, OK? If you have three to four minutes for both . . . parts . . . divide your time and if some . . . if one of you thinks that 'OK, we are talking too much for the first part' . . . change . . . go to the next question immediately, OK?

(T3, L6)

Teacher 3 advised her students to behave appropriately and maintain eye contact with their fellow candidate:

Teacher 3: OK . . . the examiner doesn't want you to look at him or her, OK? . . . because it's a conversation which will have to happen between you and the other person . . . So when you will have a picture in front of you make sure that you look at the picture and you look at the person who you are talking to . . . do not look at the examiner . . . OK . . . the examiner will give you the impression that he is not listening to you . . . OK . . . and a lot of times at the end of your discussion he or she might ask 'So what did you agree on?'

(T3, L4)

Teacher 3 also instructed her students to talk clearly and use their normal accent:

Teacher 3: We should speak clearly and slowly and remember. . . Someone from another country is not used to your accent so you have to talk clearly. Clearly! Not with a fake accent!

(T3, L5)

In addition, Teacher 3 played a video extract based on an oral examination to raise students' awareness of the body language used by candidates during the oral examination:

Teacher 3: Yeah. OK. I want you to tell me something about their gestures, about their body. What does their body language tell you?

(T3, L5)

Teacher 3 offered students various 'language phrases' to use when expressing their views on topics under discussion (Part 3):

Teacher 3: . . . How can we state our opinion? Using the phrases 'It's my belief that', 'For my part', 'As I see it', . . . or 'I . . . I'm of the opinion that', 'To my mind', 'To my way of thinking'. OK? So there are a lot of phrases which can help us express our opinion . . . 'Yes, that's right, however, I believe'. So you agree with your partner but you add another point. OK? You say something else. 'I understand what you're saying, but don't you think . . .' *What are we doing here? With this?* 'I understand what you're saying, but don't you think. . .?'

(T3, L5)

On another occasion, Teacher 3 advised her students to do extra practice on 'topics' at home:

Teacher 3: So . . . At home you can practise and you can also try to study and learn the language about the topic I gave you.

(T3, L5)

She provided her students with an interesting 'trick' to use in case they did not feel ready to initiate discussion during the oral examination:

Teacher 3: *First of all if you don't feel ready to start, you can ask the other candidate to do so, which is a devious trick. You can say 'Would you like to start?' and put him/her in the spot so that he/she cannot say 'No! You start'.*

(T3, L5)

## Summary of the findings

Through the analysis of the data presented above, it is evident that the *FCE* exam had an influence on teachers' instructional practices in the preparation programmes observed. It is important to note that the observations were conducted very early in the exam preparation cycle when exam-oriented preparation is generally expected to be less overt. As a result the intensity of influence by the *FCE* on classroom activities and discourse is particularly notable. Teachers used exam-oriented methodology to meet the requirements of the exam quite early in their effort to successfully prepare their students. For example, they focused on giving information, advice and exam tips to students as well as test-taking techniques and recommended various approaches that they felt would help their students do well in the exam. The fact that the content of the lessons resembled the *FCE* content serves as evidence that the exam had an impact on teaching.

More specifically, the exam had an impact on the reading task types used in the classroom. These resembled the types included in the *FCE* Reading paper. Other than providing students with information about the format of the reading tasks, teachers presented students with test-taking advice too. Emphasis was paid on working with unknown words in the texts, e.g. guess new words through context, ignore them or look them up in the dictionary or companion books. Teachers also focused students' attention on various ways of approaching the reading tasks, e.g. read the questions first and then the text, or vice versa. Since learning styles vary, teachers urged students to choose the approach they considered most useful in order to maximise their potential of success in the exam.

Attention was devoted to writing, too. Teachers worked on tasks that were tested in the exam such as letters, emails and narratives. Other types of writing were neglected because they were not part of the exam. Specific tenses (e.g. past tenses) were considered useful as these were deemed necessary in the narrative tasks of the exam. With regard to exam information and test-taking techniques, students were reminded to follow the word limit set by the exam and the rubrics of the writing tasks. Very often teachers presented students with model compositions that reflected the writing style and requirements of the exam. Teachers and students spent class time analysing these models. The content and layout of the model compositions were analysed in class and students spent time working on language required for each part of the model composition and its organisation depending on the writing genre. Since model compositions were used as a guide there was not much time left for original thought and creativity. The goal of the teachers was clear: students needed to become aware of the requirements of the Writing paper and practise writing tasks included in the *FCE* exam.

An equal amount of emphasis was placed on the Use of English paper, too. Extensive exposure to and practice of particular language structures and the provision of advice on the cloze tasks clearly indicate that specific features of the exam became part and parcel of the lessons observed. For example, phrasal verbs, considered very important for success in the exam, was a prominent part of students' preparation. Teachers also informed students that when working with a cloze they should be looking for collocations, phrasal verbs and auxiliary verbs rather than other types of lexis. Students were also advised to read the cloze text first, familiarise themselves with the content of the text and then attempt to fill in the gaps.

Listening was given emphasis, too. Teachers devoted time to raising students' awareness of Listening test features and proposed various ways of approaching each task so that students could choose the one that suited them best. All advice and techniques provided aimed at maximising students' test performance, which is taken as evidence of exam washback on the teaching of listening in the classes observed.

Speaking received equal attention. Teachers offered a lot of advice to their students on language to be used in the oral exam. Students were advised to speak clearly, use as many linking expressions as they could (e.g. conversation fillers, phrases for initiating and ending a conversation) and avoid repetition. Students were assisted in developing speaking skills that were directly linked to the exam such as comparing and contrasting pictures and talking in pairs in order to arrive at a negotiated decision. Students were given speaking models to follow, listened to samples of speaking interaction similar to the exam, watched a video extract of an oral examination and were asked to pay attention to body language as this was considered important for the exam.

All of the above serve as evidence of the presence of *FCE* washback on teaching. However, the question now is whether washback observed in the *FCE* preparation classes was positive or negative. The findings of the study point to the presence of both positive and negative washback. Positive washback is evident in the amount of work done on all language skills including listening and speaking plus grammar and vocabulary. Teachers spent a large amount of time on developing these skills. This is taken as an indication of positive impact of the exam as teachers, in their effort to maximise performance, worked extensively on all skills. If the exam did not include all four skills, it is doubtful whether practice on all skills would have taken place.

However, negative effects of the *FCE* test were evident on the work done on reading and writing. The fact that reading activities focused on selected-response types (e.g. multiple-choice tasks, multiple matching, etc.) emphasises a limited approach to the teaching of reading, depriving students of the opportunity to develop their skills and become critical readers through more open-ended types of reading questions. Similarly, students' writing skills were not fully developed since writing activities focused on a limited variety of genres, mainly the ones tested in the exam. The fact that students' writing was based on model compositions limited students' opportunities to develop creative writing skills.

Negative washback was observed on the teaching of listening, too. The fact that teachers trained their students to develop certain strategies to cope with the listening part could result in test-wise students who might not be able to use their listening skills for real communication purposes since the listening practice observed strictly followed the requirements of the exam. The same negative influence of the exam was seen on speaking as well since during speaking practice teachers constantly reminded their students of specific ways to respond in order to do well in the exam, instead of being encouraged to use language for authentic oral communication.

## Discussion of findings

All the above did not happen in a vacuum. Teachers were actively involved in the process of washback as they mediated between the test and the students (Tsagari 2009). Even though teachers, overall, were sensitive to students' needs for exam success and worked on enhancing students' feelings of self-efficacy, what Spratt (2005) calls 'stress inoculation', they were also actively involved in inducing washback.

Teachers tried to operationalise the exam specifications into practical, exam-oriented language activities and develop students' test-taking strategies in order to meet the needs of their prospective candidates. However, this does not necessarily mean that exam requirements were fully covered by teachers. As evidenced in the data analysed, teachers were not adequately trained in the provision of exam support. They used a limited range of techniques and were not always aware of the advice and variety of practical support provided in the *FCE Handbook for Teachers*[5], e.g. overemphasised grammar, frequently resorted to L1 to provide explanations and advice, used L2 inappropriately at times, and provided students with 'questionable' tips, especially for speaking. Their language learning theory, evidenced in the continuous provision of specific language phrases, overemphasis on phrasal verbs and other grammar areas seemed to be that language learning is made up of a series of set phrases and language skills that, if used appropriately, would lead to success in the exam. Such approaches are reminiscent of the early grammar-translation and behavouristic approaches to language teaching/learning. Nevertheless, there were differences among teachers as the type and amount of washback on teaching methods varied from teacher to teacher. For example, some teachers focused their teaching more on the exam than others: some adopted an overt 'teaching to the test' methodology (e.g. Teacher 3) while others followed more creative and independent approaches (e.g. Teacher 1).

It seems, therefore, reasonable to conclude, given the amount of variation among teachers and overall teacher practices seen so far, that the *FCE* preparation programmes are not entirely based on the exam specifications *per se*. Instead, the *FCE* programmes, and other exam-oriented programmes for that matter, are likely to be moulded according to teachers' understanding and interpretation of the exam rationale, philosophy and practical requirements. Teachers' perceptions of language teaching and learning, as well as their views on what constitutes appropriate exam preparation are likely to influence the type of teaching and learning that takes place in exam classes. Also teacher attitude towards the exam as well as the stakes and usefulness of the exam play an important role in determining the choice of methods used to teach exam classes. In the interviews teachers stressed that the *FCE* exam is a well-known exam in the present context but is considered to be difficult for the local candidates. Its wide recognition and level of difficulty are determining factors in leading teachers to resort to exam-oriented methods in their teaching approach.

Another set of factors relates to teachers' education and training in exam preparation, e.g. their personal educational experience, general and exam training, access to and familiarity with exam support materials, e.g. *FCE Handbook for Teachers*, and finally their willingness to use 'innovative' approaches to teaching and exam preparation. These teacher factors seem to have an impact on the way teachers teach exam classes.

So far, various teacher-related factors seem to have affected 'why' and 'how' teachers worked towards the exam in the present context. However, teachers, like everyone else, are part of a wider ideological, historical, economic and political context that affects their attitudes, beliefs and behaviours. *Leontief*'s theory (1978) supports that 'the motivation behind an activity could be a culturally constructed need'. Needs become motives once they are directed at a specific object. Seen in this light, the 'activity' of learning a second language in *FCE* classes in Cyprus is motivated by the culturally constructed need to obtain the certificate in order to officially establish the level of English proficiency which will consequently lead to better job opportunities or university entry. Driven by this need, students are enrolled in *FCE* classes in 'frontistiria' every year and are trained by their language teachers for the specific exam. The teachers are inevitably affected by their students' need to obtain a diploma and they are 'forced' to adapt teaching-to-the-test practices in order to help their students achieve their exam goals. The extent to which this kind of teaching helps students in achieving the desired results in exams is not yet known. However, what is known is that students who follow an exam-preparation course are likely to be influenced by teachers' approaches to exam preparation as much as by the exam itself. If teachers' understanding and practice for an exam represent a limited focus of exam specifications for students, e.g. if teachers under-represent the exam requirements, this will do a great disservice to students. In the same vein, over-emphasis on exam features is likely to give students the impression that what matters most in language learning is the focus on exam skills at the expense of other, equally relevant language skills and aspects. Therefore, teachers assume a considerable amount of responsibility for the structuring of class time, classroom interaction, and language learning. Teachers themselves, as much as the exam and exam skills, are equally responsible for what happens in the exam-preparation programmes and are an important, if not equally determining, variable of the type of teaching, and consequently learning, that takes place in exam-oriented programmes. Teachers, therefore, play a pivotal role in determining whether washback occurs, how and to what degree, which is also in line with previous research.

Finally, even though not directly researched, the textbook materials used (see also Tsagari 2009), the school – in this case the culture of 'frontistiria' – the learning traditions and the extent to which pressure is put on teachers and students for success results are likely to have an impact on instructional practices, too.

[5] http://www.lttc.ntu.edu.tw/Cambridge/MS/Handbook/FCE/fce_hb_dec08.pdf

## Conclusions and recommendations

The washback study presented in this article was a result of classroom observations of *FCE* preparation classes in three 'frontistiria' in Cyprus. The results of the study suggested that the exam, as much as teachers themselves, exercises influence on teaching and, consequently, language learning taking place within the local sociocultural context.

There were certain limitations in the implementation of this study. Lack of follow-up data was a result of time constraints. For example, the findings of the study could have been supplemented with teachers' post-observational interviews, which would have shed more light on the teaching practices observed. Although the data used in this study was adequate, collection of data from teachers for a longer period of time would give the opportunity to reveal trends that could not be observed in this study. The fact that the data was collected from a specific area (Nicosia) limited the study from investigating the scope of washback effect of the *FCE* in other areas of Cyprus. However, these results are important as they add to previous findings and, therefore, build into our understanding of the washback effect of language tests.

The results of the study have implications for exam-preparation programmes. Teachers of exam-preparation classes, for instance, need to be well aware of the exam aims and specifications in order to prepare their students efficiently. Such teachers should appropriately familiarise students with exam requirements. However, teachers need to keep a balance between language teaching/learning and preparation for the exam and imbue their classes with communicatively oriented language opportunities designed with clear learning aims and objectives that do not relate to the requirements of the exam only. Teachers, with appropriate training, can exploit exam preparation for language learning maximising students' learning potential.

Exam constructors could play a crucial role in changing teachers' perception and approach towards the exam, e.g. keep exams up to date with the current learning and language learning theories and constantly revise them. The exam specifications included in exam handbooks should clearly state the intended washback of the exam along with highly practical recommendations about how to achieve this. Since teachers play an important role in the washback process, appropriate communication channels need to be established and maintained. Examination bodies should provide detailed guidelines and detailed feedback on exam results to teachers and make sure that exam washback is beneficial for both teaching and learning.[6]

Researchers wishing to investigate test washback need to have a detailed and clear course of action for their research study, e.g. a clear understanding of the aim of their investigation and appropriate methodology. Furthermore, future researchers need to return to the teachers after the data collection and analysis of the lesson transcripts to investigate the reasons behind teachers' actions and teaching practices. Future research should also focus on test preparation effects on test performance and English language proficiency and research the degree of intensity of exam washback as the dates of the exam are drawing nearer (see also Tsagari 2009).

The results of the study reported here revealed the complexity of high-stakes exam preparation which, according to sociocultural theory and research, engages students in multi-dimensional social practices where instruction constitutes a process of socialisation into test-taking behaviours and into the priorities embodied in the exam. The study has also provided evidence of the consequential validity of the *FCE* exam and produced descriptions of classroom practice and characteristics of the teachers and other factors that facilitate or hinder positive impact. Finally, given that no other study on *FCE* washback in Cyprus was conducted before, it is hoped that the present study be used as a basis for further research into the washback effect of *FCE* and other exams in the Cypriot context.

## References and further reading

Alderson, J C and Hamp-Lyons, L (1996) TOEFL preparation courses: A study of washback, *Language Testing* 13 (3), 280–297.

Alderson, J C and Wall, D (1993) Does washback exist? *Applied Linguistics* 14 (2), 115–129.

Bailey, K M (1996) Working for washback: A review of the washback concept in language testing. *Language Testing* 13 (3), 257–279.

Bailey, K M (1999) *Washback in Language Testing*, TOEFL Monograph Series. Report Number: RM-99-04, TOEFL-MS-15. Princeton, NJ: Educational Testing Service. Retrieved 11/3/2001, from http://www.ets.org

Becker, B (1990) Coaching for the SAT: Further synthesis and appraisal, *Review of Educational Research* 60, 373–417.

Brindley, G and Ross, S (2001) EAP assessment: issues, models, and outcomes, in Flowerdew, J and Peacock, M (Eds) *Research Perspectives on English for Academic Purposes*, Cambridge: Cambridge University Press, 148–166.

Brown, H D (2004) *Language assessment: Principles and classroom practices*, London: Longman.

Burrows, C (2004) Washback in Classroom-based Assessment: A Study of the Washback Effect in the Australian Adult Migrant English Program, in Cheng, L, Watanabe, Y and Curtis, A (Eds) *Washback in Language Testing: Research Context and Methods*, Mahwah, NJ: Lawrence Erlbaum Associates, Inc, 113-128.

Cheng, L (2005) *Changing Language Teaching Through Language Testing: A washback study*, Cambridge: Cambridge University Press.

Cohen, L, Manion, L and Morrison, K (2000) *Research Methods in Education* (5th ed.), New York: Routledge Falmer.

Elder, C and O' Loughlin, K (2003) Investigating the relationship between intensive English language study and band score gain on IELTS, *British Council/IDP Australia IELTS Research Reports* volume 4, 208–254. Available online from http://www.ielts.org/researchers/research.aspx

Green, A (2003) *Test impact and English for academic purposes: a comparative study in backwash between IELTS preparation and university pre-sessional courses*, unpublished PhD thesis, Centre for Research in Testing, Evaluation and Curriculum in ELT, University of Surrey, Roehampton, UK.

Green, A (2006) Washback to the learner: learner and teacher perspectives on IELTS preparation course expectations and outcomes, *Assessing Writing* 11, 113–134.

[6] Such has been the recent practice of Cambridge ESOL, see http://www.cambridgeesol.org/index.html

Hayes, B and Read, J (2004) IELTS Test Preparation in New Zealand: Preparing Students for the IELTS Academic Module, in Cheng, L, Watanabe, Y and Curtis, A (Eds) *Washback in Language Testing: Research Context and Methods*, Mahwah, NJ: Lawrence Erlbaum Associates, Inc, 97–112.

Kramsch, C J (Ed.) (2002) *Language acquisition and language socialization: ecological perspectives*, London/New York: Continuum.

Lantolf, J P (Ed.) (2000) *Sociocultural Theory and Second Language Learning*, Oxford: Oxford University Press.

Lantolf, J P and Thorne, S L (2006) *Sociocultural Theory and the Genesis of Second Language Development*, Oxford: Oxford University Press.

Messick, S (1996) Validity and washback in language testing, *Language Testing* 13, 241–256.

Mickan, P (2000) *Textualising meanings: second language writers in action*, unpublished PhD thesis, Macquarie University, Australia.

Mickan, P (2003) What's your score? An investigation into language descriptors for rating written performance, *British Council/IDP Australia IELTS Research Reports* volume 5, 125–157. Available online from http://www.ielts.org/researchers/research.aspx

Mickan, P (2006a) Socialisation through teacher talk in an Australian bilingual class, *International Journal of Bilingual Education and Bilingualism* 9 (3), 342–358.

Mickan, P (2006b) Socialisation, social practices and teaching, in Mickan, P, Petrescu, I and Timoney, J (Eds) *Social practices, pedagogy and language use: studies in socialisation*, Adelaide: Lythrum Press, 7–23.

Mickan, P and Motteram, J (2008) An ethnographic study of classroom instruction in an IELTS preparation program, *British Council/IDP Australia IELTS Research Reports* volume 8, 17–43. Available online from http://www.ielts.org/researchers/research.aspx

Mickan, P and Slater, S (2003) Text analysis and the assessment of academic writing, *British Council/IDP Australia IELTS Research Reports* volume 4, 59–88. Available online from http://www.ielts.org/researchers/research.aspx

Mickan, P, Slater, S and Gibson, C (2000) A study of response validity of the IELTS Writing subtest, *British Council/IDP Australia IELTS Research Reports* volume 3, 29–48. Available online from http://www.ielts.org/researchers/research.aspx

Muhr, T (2004) *User's Manual for ATLAS.ti 5.0* (2 ed.) Berlin: Scientific Software Development.

Munoz, A P and Alvarez, M E (2010) Washback of an oral assessment system in the EFL classroom, *Language Testing* 27 (1), 33–49.

Qi, L (2005) Stakeholders' conflicting aims undermine the washback function of a high-stakes test, *Language Testing* 22 (2), 142–173.

Rao, C, McPherson, K, Chand, R and Khan, V (2003) Assessing the impact of IELTS preparation programs on candidates' performance on the General Training Reading and Writing test modules, *British Council/IDP Australia IELTS Research Reports* volume 5, 237–262. Available online from http://www.ielts.org/researchers/research.aspx

Robb, T N and Ercanbrack, J (1999) A study of the effect of direct test preparation on the TOEIC scores of Japanese university students, *TESL-EJ* 3 (4). Retrieved from http://tesl-ej.org/ej12/toc.html

Saif, S (2006) Aiming for positive washback: a case study of international teaching assistants, *Language Testing* 23 (1), 1–34.

Shohamy, E (1993) The power of test: The impact of language testing on teaching and learning, *National Foreign Language Center Occasional Papers*, Washington, DC: National Foreign Language Center.

Shohamy, E, Donitsa-Schmidt, S and Ferman, I (1996) Test impact revisited: washback effect over time, *Language Testing* 13 (3), 298–317.

Spratt, M (2005) Washback and the classroom: the implications for teaching and learning of studies of washback from exams, *Language Teaching Research* 9 (1), 5–29.

Stecher, B, Chun, T and Barron, S (2004) The Effects of Assessment-Driven Reform on the Teaching of Writing in Washington State, in Cheng, L, Watanabe, Y and Curtis, A (Eds) *Washback in Language Testing: Research Context and Methods*, Mahwah, NJ: Lawrence Erlbaum Associates, Inc, 53–72.

Swain, M, Kinnear, P and Steinman, L (2010) *Sociocultural Theory in Second Language Education: An Introduction Through Narratives*, Bristol, UK: Multilingual Matters Ltd.

Torrance, H (Ed.) (1995) *Evaluating authentic assessment*, Buckingham: Open University Press.

Tsagari, D (2009) *The Complexity of Test Washback: An Empirical Study* (volume 15), Frankfurt am Main: Peter Lang GmbH.

UCLES (2007) *First Certificate in English: Handbook for Teachers for Examinations from 2008 onwards*, University of Cambridge. ESOL Examinations. Available online https://www.teachers.cambridgeesol.org/ts/digitalAssets/109701_fce_hb_dec08.pdf

Unsworth, L (Ed.) (2000) *Researching language in schools and communities: functional linguistic perspectives*, London: Cassell.

Van Lier, L (1988) *The Classroom and the Language Learner: Ethnography and Second Language Classroom Research*, London: Longman.

Wall, D (2005) *The Impact of High-Stakes Examinations on Classroom Teaching: A case study using insights from testing and innovation theory*, Cambridge: Cambridge University Press.

Wall, D and Alderson, J C (1993) Examining washback: The Sri Lankan impact study, *Language Testing* 10 (1), 41–69.

Watanabe, Y (1997) *The Washback Effects of the Japanese University Entrance Examinations of English-classroom-based Research*, unpublished PhD thesis, Department of Linguistics and Modern English Language, Lancaster University, UK.

Weeden, P, Winter, J and Broadfoot, P (2002) *Assessment: what's in it for schools?* London: Routledge.

Wesdorp, H (1983) Backwash effects of multiple-choice language tests: myth or reality? in van Weeren, J (Ed.) *Practice and problems in language testing 5: non-classical test theory/final examinations in secondary schools*, Arnhem: Cito, 85–103.

# ALTE briefing

## ALTE's 40th meeting and conference

ALTE's 40th meeting and conference took place in Bochum, Germany, 16–18 November, and was hosted by one of ALTE's three German members, TestDaF-Institut. The first two days included a number of workshops and Special Interest Group meetings for ALTE members and affiliates, and the third day was an open conference day for all those with an interest in language testing.

The theme of the conference was *Achieving Context Validity* and Professor Gillian Wigglesworth, Professor of Linguistics and Applied Linguistics at the University of Melbourne ran a workshop and gave a plenary presentation on task-based testing. Professor Cyril Weir, Powdrill Professor in English Language Acquisition at the University of Bedfordshire and Director of the Centre for Research in English Language Learning and Assessment ran a workshop and gave a plenary presentation on issues related to context validity parameters.

There were also plenary presentations from Professor Günther Sigott, Associate Professor in Applied Linguistics at the University of Klagenfurt, Austria, and Dr Evelina Galaczi, Senior Research and Validation Manager at Cambridge ESOL, as well as from Dr Thomas Eckes, Head of the Psychometrics and Research Methodology Department at TestDaF-Institut and Sonja Zimmermann, a Test Development Officer at TestDaF-Institut. In addition, Michael Corrigan from the ALTE Validation Unit ran a workshop with several ALTE colleagues – Siuán Ní Mhaonaigh of the Language Centre, NUI Maynooth, Danilo Rini from CVCL, University for Foreigners, Perugia, and Paul Crump, Senior Assessment Manager at Cambridge ESOL.

Following the conference, ALTE ran a two-day Introduction to Assessing Speaking Course run by Dr Evelina Galaczi and Lucy Chambers from the Research and Validation Group at Cambridge ESOL, and Annie Broadhead, Consultant to Cambridge ESOL ran a one-day Foundation Course in Language Testing: Getting Started. Both courses were well attended with participants coming from several countries, within Europe and beyond.

## ALTE's 41st meeting and conference

Registration is already underway for ALTE's next meeting and conference, which will take place in Lisbon, 18–20 April 2012. The theme of the conference is *The Impact of Language Testing on Learning and Teaching* and keynote speakers will include Professor Norbert Schmitt of the University of Nottingham, Dr Dianne Wall of the University of Lancaster and Trinity College, London, Dr Nick Saville of University of of Cambridge ESOL Examinations, and Professor Maria José Grosso of the Centre for the Evaluation of Portuguese as a Foreign Language at the University of Lisbon. The conference will be preceded by ALTE Auditee/Auditor Orientation and Training, and an ALTE Introductory Testing Course, 16–17 April, and will be followed by an ALTE Foundation Course on 21 April.

For further information about all ALTE activities, please visit the ALTE website – www.alte.org. To become an Individual Affiliate of ALTE, please download an application form from the ALTE website or contact the Secretariat – info@alte.org. Individual affiliation to ALTE is free of charge and means you will receive advance information of ALTE events and activities and an invitation to join the ALTE electronic discussion forum.

# Caroline Clapham IELTS Masters Award 2011

Since 2000, the IELTS partners have presented the Caroline Clapham IELTS Masters Award to the Master's-level dissertation or thesis in English which makes the most significant contribution to the field of language testing.

Recently, the IELTS Research Committee announced the selection of Kellie Frost as the winner of the 2011 award. Her dissertation investigated the validity of an integrated listening–speaking task using an innovative discourse analysis-based methodology. Multiple reviewers praised it for its clear formulation of research questions, good rationale for choosing the unit of analysis, astute interpretations of data, skilful use of references, and for its contribution to the field conceptually and methodologically. The dissertation was submitted to the University of Melbourne and was supervised by Prof Catherine Elder. The abstract for the dissertation appears below.

Kellie will be presented with her award – a certificate and a cheque for £1,000 – at the 2012 Language Testing Research Colloquium in Princeton, New Jersey. Qualified individuals who would like to join the 2012 competition are invited to visit http://ielts.org/researchers/grants_and_awards/ielts_masters_award.aspx for details of the competition and submission guidelines.

### Investigating the validity of an integrated listening–speaking task: A discourse-based analysis of test takers' oral performances

**Ms Kellie Frost, University of Melbourne**

Performance on integrated tasks requires candidates to engage skills and strategies beyond language proficiency alone, in ways that can be difficult to define and measure for testing purposes. While it has been widely recognised that stimulus materials impact test performance, our understanding of the way in which test takers make use of these materials in their responses, particularly in the context of listening–speaking tasks, remains predominantly intuitive, with little or no base in empirical evidence.

Limited studies to date on integrated Speaking tests have highlighted the problems associated with content-related aspects of task fulfilment (Brown et al (2005) *TOEFL Monograph Series MS-29*; Lee (2006) *Language Testing* 23: 131), but little attempt has been made to operationalise the way in which content from the input material is integrated into speaking performances. Using discourse data from a trial administration of the new Oxford English Language Test, this thesis investigates the way in which test takers integrate stimulus materials into their speaking performances on an integrated listening then speaking summary task, and examines if test scores reflect real differences in the quality of oral summaries produced.

The study will address the following validity issues: Firstly, whether the integrated listening–speaking task is measuring a construct of speaking ability common to the other speaking tasks; secondly, if speaking ability (as measured by the overall speaking score on the test) corresponds to real differences in the content quantity and quality of oral performances on the integrated listening–speaking task; and finally, if the discourse produced by test takers in response to this integrated task provides empirical support for the task rating scale descriptors. An innovative discourse analytic approach was developed to analyse content-related aspects of performance in order to determine if such aspects represent an appropriate measure of the construct of speaking ability. Results showed that the quantity and quality measures devised to operationalise content, such as the number of key points included from the input text, and the accuracy with which information from the input text was reproduced or reformulated, effectively distinguished participants according to their level of speaking proficiency, indicating that these discourse-based measures have the potential to be applied to other integrated tasks and in other assessment contexts.

# Winner of the 2012 Cambridge/ILTA Lifetime Achievement Award

The winner of the 2012 Cambridge/ILTA Lifetime Achievement Award was announced by the Chair of the 2012 Award Committee, Elana Shohamy, to the language testing community at the end of 2011. Here we present her citation of the winner, Professor Carol Chapelle of Iowa State University, USA.

On behalf of the committee for the 2012 Cambridge/ILTA Lifetime Achievement Award (consisting of John Read, Jo Lewkowicz, Hanan Khalifa, and myself), it is a great pleasure to announce that, after carefully considering the several highly meritorious nominations received from the field, the committee has selected Professor Carol Chapelle to receive the Lifetime Achievement Award to be presented at the 34th Language Testing Research Colloquium, which will be held in Princeton, USA in April 2012.

Professor Carol Chapelle is a Distinguished Professor in TESL/Applied Linguistics in the Department of English of Iowa State University. Throughout her career she has undertaken a program of research and publications in language testing which has been deeply embedded in a wider range of interests in applied linguistics and TESOL. She has made notable contributions in two major areas of the field. The first is in the use of computer technology for language testing, growing originally out of her experience with computer-assisted language learning (CALL) as an ESL teacher. Since then, she has been a leader in investigating the potential of the technology to enhance language assessment, while at the same time maintaining a critical perspective by acknowledging problem areas and challenges. She is the author (with Dan Douglas) of *Assessing Language Through Computer Technology* (Cambridge, 2006), a comprehensive survey of the area. More broadly, Carol's work on computer-based assessment should be viewed as a key component of her primary interest in issues at the intersection of computer technology and applied linguistics, as reflected in her books *Computer Applications in Second Language Acquisition* (Cambridge, 2001) and *English Language Learning and Technology* (Benjamins, 2003).

The second area in which Carol has made outstanding contributions is the construct validation of language tests. Through a series of very influential papers in the late 1990s and early 2000s, she explored how modern validity theory could be applied in the analysis and development of language tests, particularly but not exclusively those designed to assess vocabulary knowledge and ability. At the same time she was deeply involved in building the conceptual framework for what has become the internet-based Test of English as a Foreign Language (iBT). This was the basis for  a sophisticated validity argument, as presented in the volume for which she was first co-editor and a prominent author, *Building a Validity Argument for the Test of English as a Foreign Language* (Routledge, 2008) – a book described by Alister Cumming in a *Language Testing* review as a "monumental achievement".

Apart from her theoretical contributions, Carol was co-director with Joan Jamieson of the project that led to the Longman English Assessment and has developed innovative language tests for her own institution.  She was also co-author of *ESOL Tests and Testing: A Resource for Teachers andAdministrators* (TESOL, 2005), a noteworthy initiative to promote assessment literacy among the target readership.

Carol has served on the Executive Board of ILTA and been a frequent presenter at LTRC, as well as an active member of the Midwest affiliate MwALT. She won the ILTA Best Article Award in 1998 for her book chapter "Construct definition and validity inquiry in SLA research".  She has also served the wider field as Editor of *TESOL Quarterly* and President of AAAL – roles which have brought her to international prominence and provided opportunities to communicate her work on language testing and related areas to a much broader audience. Her current major project is the multi-volume Wiley-Blackwell *Encyclopedia of Applied Linguistics*, for which she is not only General Editor but also editor of the Assessment volume.

Because of her many significant contributions over the years, the committee is delighted to select Professor Carol Chapelle to receive the 2012 Cambridge/ILTA Lifetime Achievement Award.

# Studies in Language Testing

Volume 30 from the *Studies in Language Testing* series was published in September 2011. The volume, edited by Lynda Taylor, is entitled *Examining Speaking: Research and Practice in Assessing Second Language Speaking.*

This volume develops a theoretical framework for validating tests of second language speaking ability. The framework is then applied through an examination of the tasks in Cambridge ESOL speaking tests from a number of different validity perspectives that reflect the socio-cognitive nature of any assessment event. The chapter authors show how an understanding and analysis of the framework and its components can assist test developers to operationalise their speaking tests more effectively, especially in relation to the key criteria that differentiate one proficiency level from another.

This volume is a rich source of information on all aspects of examining speaking ability. It provides an up-to-date review of the relevant literature on assessing speaking, an accessible and systematic description of the different proficiency levels in second language speaking, and a comprehensive and coherent basis for validating tests of speaking.

This volume will be of considerable interest to examination boards and other test providers who wish to validate their own speaking tests in a systematic and coherent manner, as well as to academic researchers and graduate students in the field of language assessment more generally. This is a companion volume to the previously published titles *Examining Writing* and *Examining Reading*.

Information on all the volumes published in the SiLT series is available at: http://research.cambridgeesol.org/research-collaboration/silt

Research

The URL for reading/downloading issues of *Research Notes* is:
**www.CambridgeESOL.org/research-notes**

The URL for subscribing to *Research Notes* is:
**www.CambridgeESOL.org/join-research-notes**

# Contents:

For further information visit the website:
**www.CambridgeESOL.org**

University of Cambridge
ESOL Examinations
1 Hills Road
Cambridge
CB1 2EU
United Kingdom
Tel. +44 1223 553997
Email ESOLhelpdesk@CambridgeESOL.org

ALTE Association of Language Testers in Europe

PART OF THE
CAMBRIDGE ASSESSMENT
GROUP