Research Notes

Contents

Editorial Notes

Welcome to issue 30 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL.

In this issue we focus on the processes and outcomes involved in reviewing our exams, with specific reference to the review of the First Certificate in English (FCE) and Certificate of Advanced English (CAE) which started in 2004 and will culminate in December 2008 with the first administration of the updated exams. Cambridge ESOL carries out regular reviews of our assessment products to ensure that they remain accessible and relevant to the changing demands of stakeholders and developments in the language testing field. This issue provides an overview of the FCE and CAE Review Project and presents a range of major research and consultation activities - together with their outcomes - undertaken both within this project and for other exams.

In the opening article, Roger Hawkey gives an overview of the FCE and CAE Review Project, providing its historical context, previous revisions and updates, and some of the major themes which informed the review. Ardeshir Geranpayeh then reports on studies undertaken to investigate how Structural Equation Modelling (SEM) can aid the revision of high stakes testing, using CAE as a case study. He describes how SEM can be used to show that changes to the format of tests would not significantly change the underlying constructs of the CAE exam.

We then include articles on the four major skills papers, namely Reading, Writing, Listening and Speaking. Firstly, Helen Coward summarises research on the inclusion of short themed texts in the CAE Reading paper. Next, Margaret Cooze and Stuart Shaw report on research to establish the impact of reduced input and output length in FCE and CAE Writing papers using a series of multiple rating exercises where groups of examiners rated common sets of writing performances for the updated specifications.

The next two articles focus on Listening. Steve Murray describes research on the CAE Listening test which trialled the changing of alternative tasks within it to a fixed format. This is followed by a summary by Diana Fried-Booth of research into FCE Listening focusing on changes to the format of Part 1 of the paper. Turning to Speaking, Clare Harrison reports on research and consultation undertaken to review the FCE and CAE Speaking tests. Widening the focus to include Business as well as General English, Evelina Galaczi and Angela ffrench describe the revision of assessment scales for Speaking tests for Main Suite and Business English Certificate (BEC) exams.

Finally, Fiona Barker, Steve Murray, Stephen McKenna and Ivana Vidakovic outline a range of other research and stakeholder projects undertaken within the FCE and CAE Review Project. This is followed by conference reports and a registration call for the ALTE 2008 conference we are hosting in Cambridge.

Editorial team for Issue 30: Fiona Barker, Anne Gutch, Angela ffrench and Kirsty Sylvester.

The 2004–2008 FCE and CAE Review Project: historical context and perennial themes

ROGER HAWKEY ESOL CONSULTANT

Introduction

Aldous Huxley (1952) considers the 'charm of history and its enigmatic lesson consist in the fact that, from age to age, nothing changes and yet everything is completely different'. He would thus probably have favoured a view of the 2004 to 2008 First Certificate in English and the Certificate in Advanced English (FCE and CAE) Review Project in its historical context. Cambridge ESOL certainly sees the sense of an historical approach. See, for example, recent publications in the Studies in Language Testing (SiLT) series: Volume 15, *Continuity and Innovation: Revising the Cambridge Proficiency in English examination 1913–2002* (Weir and Milanovic 2003); Volume 16, *A Modular Approach to Testing English Language Skills* (Hawkey 2004), which traces the development of the Certificates in English Language Skills (CELS) from predecessor exams, and Volume 23, *Assessing Academic English: Testing English proficiency 1950–1989 – the IELTS solution* (Davies in press). Further back, the very first volume in the SiLT series, *An investigation into the comparability of two tests of EFL* (Bachman, Davidson, Ryan and Choi 1995), took a hard look at FCE and TOEFL before a major revision of FCE in 1996.

In this tradition, a further SiLT volume is due for publication around the time of the updated versions of the FCE and CAE exams at the end of 2008. This will trace the histories of both exams and conclude with an account of the FCE and CAE Review Project. Reference to this volume is made here to set the context for other articles in this issue of *Research Notes* with its focus on FCE and CAE. This article will take a quick journey through the histories of the two exams, with particular reference to their past revisions and common themes influencing them, then summarise the current review.

Historical context and previous exam revisions

The FCE started life as the Lower Certificate in English (LCE), in 1939. By then international demand was growing for an exam at a lower level than the Certificate of Proficiency in English (CPE), which had been in operation since 1913. CPE was pitched at a proficiency level then described as for candidates with an 'accurate use of idiomatic English', and 'equal to a pass with credit in English in the School Certificate Examination' (*Regulations*, UCLES, 1939). The level of the prescribed texts for the new exam was also somewhat imprecisely defined, as intended to 'provide reading matter of a suitable standard of difficulty and to form the basis of the relatively limited vocabulary – which, it is recognized, is all that can be expected at this stage' (ibid.). Here was the first recurring theme in the history of the FCE and CAE exams, then and now: the problem of defining and comparing *levels of proficiency*. In the summary of the FCE and CAE Review Project below, the mapping of the updated exams to the Common European Framework of Reference (CEFR) by 'levels of proficiency which allow learners' progress to be measured at each stage of learning' (Council of Europe 2001:1) is a key focus. There was no CEFR back in 1939, of course.

In fact, a different but related problem with levels led to quite fierce argument at the time FCE was proposed. Some key UCLES stakeholders were concerned by the very idea of an English exam designed to test the language at a 'basic' English level. This was perceived, perhaps, as in conflict with a conception of exams of English as it is spoken by native or near-native speakers of the language. Consultation and sometimes disagreement with *partners and other stakeholders* is another recurring theme in the development of UCLES EFL then Cambridge ESOL exams over the decades.

Table 1 summarises the main LCE/FCE and CAE exam revisions since 1939. The notes on each revision highlight innovation and modification, for example:

- growing systematicity in the revision of the Cambridge ESOL Upper Main Suite exams
- evolution from a literary English examination tradition towards a communicative language proficiency construct incorporating the components of language in their hierarchical relationships and in authentic task use
- a requirement of comprehensive and precise exam specifications, a third historical theme, including evidence-based and transparent quantitative and qualitative test validation processes.

The importance of our recurring themes of *language levels* and *stakeholder relations* is also implicit in the Table 1 summaries.

The FCE and CAE Review Project

The first of a series of FCE and CAE Review bulletins entitled 'Reviewing FCE and CAE' (Cambridge ESOL 2005a) proposes the model in Figure 1 for the Review Project.

This framework attempts to link stakeholder consultation, research and test trialling in the dynamic processes of exam specification and approval.

Project purpose, aims, processes

The November 2006 *Bulletin* entitled 'FCE and CAE – Exam specifications from December 2008' (Cambridge ESOL 2006c) documents the purpose, aims, outcome and process of the Review. The stated *purpose* of the project was to ensure that FCE and CAE meet the current needs of candidates, teachers, centres and other users in terms of

Table 1: Summary of LCE, FCE and CAE exams, 1939 to 2003 (continued overleaf)

Year	Exam	Summary Content	Main innovation and changes
1939	**LCE**	**Total time: 4 hrs + the Oral** *Oral:* Dictation, Reading, Conversation *Written:* (a) Prescribed texts (b) English Composition and Language	
1946	**LCE**	**Total time: 5 hrs + the Oral** *Oral:* Dictation, Reading, Conversation *Written:* (a) Translation from and into English or Prescribed Books* 2½ hours (b) English Composition and Language 2½ hours	• stabilised and more fully specified exam post World War 2 • translation now included • *Local Secretaries had the option to arrange with Cambridge for a paper on prescribed texts in place of translation.
1951	**LCE**	**Total time: 5 hrs + the Oral** *Oral:* Dictation, Reading, Conversation *Written:* (a) Prescribed Books or Translation from and into English 2½ hours (b) English Composition and Language (composition and a passage of English with language questions) 2½ hours	Written exam by now contained: • Prescribed Books *or* • Translation from and into English, *and* • English Composition and Language (with an ESP element).
1970	**LCE**	**Total time: 5 hrs 40 m** *Oral Tests:* Reading, Conversation, Listening Comprehension (40m) Composition, Comprehension (2½ hrs) + *one* of the following 2½ hr papers: (a) Translation from and into English (b) Structure and usage (c) Prescribed Books	• Dictation replaced by Listening Comprehension i.e. written answers to questions across range of sources, some responses in multiple-choice form. • Composition and Comprehension paper also has multiple-choice items, sentence completion or re-arrange formats • New Structure and Usage paper, cf.: Use of English paper in CPE exam since 1953.
1975	**FCE**	**Total time: 5 hrs 55 m** Paper 1: Composition Paper 2: Reading Comprehension Paper 3: Use of English Paper 4: Listening Comprehension Paper 5: Interview	• LCE renamed First Certificate of English (FCE) • now three passing, two failing grades • compulsory papers for uniformity in equating candidate performances across papers • translation paper available but not now affecting grades awarded • 25% of marks to Papers 1, 2 and 3 and Papers 4 and 5 combined • no more prescribed books • composition separated from “Language” and comprehension • stronger focus on assessment criteria • CPE-type *Use of English* paper in place of the LCE *Structure and Usage* • range of semi-objective and objective task formats to test vocabulary, grammatical and lexical forms, sentence structure • all multiple-choice testsconstructed from pre-tested items.
1984	**FCE**	**Total time: 5 hrs 20 m** *(Note: Paper renumbering)* Paper 1: Reading Comprehension Paper 2: Composition Paper 3: Use of English Paper 4: Listening Comprehension Paper 5: Interview	• UCLES EFL exam revisions more regular process now • strengthening communicative approach to language testing (CALT(e)) • more rigorous psychometric test validation measures • new optional composition task based on a candidate’s reading of prescribed texts, but with assessment emphasis on “control of language in the given context” • Paper 4 now uses recorded material e.g. radio news, features, situational dialogues, announcements; charts, diagrams, picture prompts; all objective items; no literature-oriented texts • Paper 5 increase in oral weighting and realism in test tasks, more role play • marks for aural/oral skills: Papers 4 and 5 now take ⅓ not ¼ of total marks.
1991	**CAE**	**Total time: 5 hrs 30 m** Paper 1: Reading Paper 2: Writing Paper 3: English in Use Paper 4: Listening Paper 5: Speaking	• CAE introduced in 1991 • level between FCE and CPE & overtly communicative construct • 20% of total score for each paper • designed to fit within PET-FCE-CPE suite • emphasis on authenticity of tests and tasks • more communicative language oriented assessment criteria • broad range of language micro-skills tested, including correcting and editing • one multi-source writing task, integrating skills • double marking of writing paper • maximum use of optically marked scan answer sheets.
1996	**FCE**	**Total time: 4 hrs 25 m** Paper 1: Reading Paper 2: Writing Paper 3: Use of English Paper 4: Listening Paper 5: Speaking	• Comparability Study (1987–1989) • new UCLES EFL Validation department • fuller, theory- and construct-based *Specifications* • *joint* FCE and CPE revision project from 1991 – most extensive consultation exercise ever • influences from new CAE exam • 1990s focus on authenticity, learners’ learning strategies • construct-based changes • increased trialling • improved IT support systems and increased staffing • writing output genres agreed after statistical analyses of trial tests: personal and transactional letters, articles, reports and compositions • Use of English paper construct: candidate knowledge of lexico-grammatical systems to convey meanings not just manipulating • “mechanics of system” validated through Rasch analysis • increasing explicit relating of UCLES exams first to ALTE 5-level system, then to Council of Europe (CoE) levels.

Table 1: Summary of LCE, FCE and CAE exams, 1939 to 2003 *(continued)*

Year	Exam	Summary Content	Main innovation and changes
1999	**CAE**	**Total time 5 hrs 30 m** Paper 1: Reading Paper 2: Writing Paper 3: English in Use Paper 4: Listening Paper 5: Speaking	• Paper 3 new word formation task • no expansion of notes into sentences task • Paper 5 no 'describe and draw' task.
2003	**FCE**	**Total time 4 hrs 25 m** Paper 1: Reading Paper 2: Writing Paper 3: Use of English Paper 4: Listening Paper 5: Speaking	• NB no substantive changes in format since 1996.

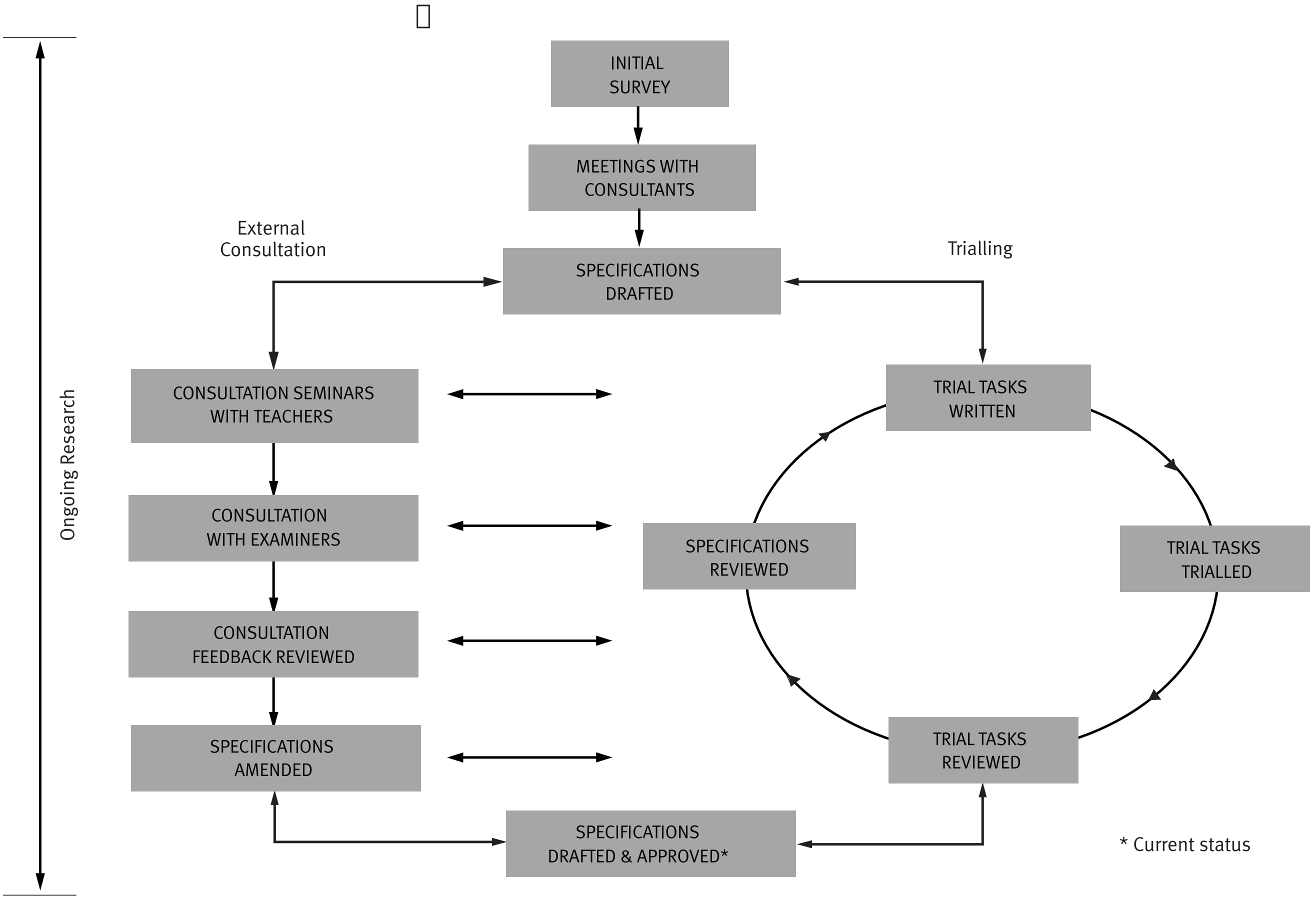
content and length. In pursuit of this purpose, the Project should reflect developments in language teaching and learning together with developments in Cambridge ESOL's other General English examinations. Taking account of information from the Candidate Information Sheets, the project should aim to achieve 'a thoroughly validated examination', defining a specific test focus for each part of each paper, to ensure that the updated exams meet the needs of candidates and other users.

The *processes* of the project would reflect the model in Figure 1, namely: data collection from key stakeholders, the development of exam specifications, the definition of the test foci and uses; the production, editing and trialling of draft task types and materials; development and trialling of assessment criteria; and research into the validity and reliability of the material and assessment procedures. Exam support materials, including training programs for examiners and writers of examination materials, must also be developed.

Review issues

As throughout the history of FCE and CAE, the decision to review exam papers was based on matters of theory and practice arising both from the broader language testing context and from continuing experience of and research into running the exams. So it was with the FCE and CAE Review Project. The May 2005 *Bulletin* (Cambridge ESOL 2005a) indicates potential modification areas already foreseen and subject to 'initial investigations'. These were expressed as four questions relevant to FCE and CAE, but 'which may apply to other exams as well in the long term' (2005:1).

Figure 1: Model for the FCE and CAE Review Project



The questions were:

1. Whether it would be desirable to *reduce the length* of the FCE/CAE exams.
2. How strong demand was for optional *computer-based versions of the exams*, and what the implications were of introducing computer-based tests.
3. Whether *enhanced certification* would be desirable, that is providing an award at the Council of Europe level below that of the exam taken for candidates who fail narrowly to achieve the passing grade in the exam they take.
4. Whether demand was strong for *optional English for Specific Purposes (ESP)* tests to supplement the four skills papers.

In addition to these four broad questions, the Review Project was to be guided by the perceived need to:

- 'ensure that the task types and their test focus continue to reflect the needs of candidates'
- 'make the most effective use of resources such as language corpora and new technology (relating to advances in the marking of tasks)'
- 'review of all the components of FCE and CAE' to keep 'up to date with changes in methodology' (Cambridge ESOL 2005a).

The experience of the past review projects, especially the 1991–1996 FCE and the 1991–2002 CPE reviews, had emphasised the value of a full-scale action schedule. The regularly updated plan for the 2004–2008 review of the FCE and CAE examinations specified no fewer than 138 action areas, categorised by their function in the review process, and their target start and finish dates. As would be expected from its extensive remit, the project design incorporated extensive inter-departmental collaboration. The monthly meetings of the Modifications to Examinations Working Group (MEWG) received regular reports of their project involvements from the Assessment and Operations Group (AOG), in particular the Upper Main Suite team, Research and Validation Group, Pretesting and Performance Testing Units, the Business Support Group and the Projects Office.

Consultation

Stakeholder consultation for the FCE and CAE Review Project was broad. The Main Suite Modification (MSM) Market Research Report on the FCE and the CAE exams (Chamberlain 2004) was based on data from questionnaires to Cambridge ESOL exam candidates, Local Secretaries, Directors of Studies, Examiners, Examinations Officers, teachers and teacher trainers, and materials writers. The geographical focus of the survey was on Cambridge ESOL's main markets in Asia/Australasia, the Benelux countries, Central and Eastern Europe, France, Germany, Greece, Italy, Latin America, the Middle East, Portugal, South Asia, Spain, Switzerland and the United Kingdom. Responses were received from 1,900 candidates and 726 other stakeholders across 20 countries.

In addition to the survey, opinions on possible FCE and CAE updates were sought through the regular meetings of the MEWG and the General English Steering Group; reports from external consultants, Chairs (professionals external to Cambridge ESOL responsible for the content of the papers) and Subject Officers of papers; skill-focused review meetings with Chairs and Item Writers, where each part of each paper was evaluated and reviewed, and invitational meetings. The invitational meeting of 4 April 2006, for example, involved: Cambridge ESOL staff, Principal Examiners, Senior Team Leaders and Local Secretaries; FCE and CAE stakeholders from the British Council, English UK, the European Association for Quality Language Services (EAQUALS), Reading University and Cambridge University Press. The historical partnership and stakeholder theme of previous decades, as can be seen, persisted strongly.

The research agenda

At the request of the MEWG, the Research and Validation Group produced, in June 2007, a *Summary of research activities associated with FCE and CAE modifications*. This report summarised research commissioned by three groups within Cambridge ESOL (Assessment and Operations, Business Support and Research and Validation Groups). The contents of the report were expected to 'feed into internal briefing documents for ESOL staff as well as into various elements of the external communications strategy associated with the modifications project (e.g. *Cambridge First*, teacher seminar materials, *Research Notes*, website pages)'.

In fact, it is not straightforward to list or count research studies for the FCE and CAE Review Project. This is because studies on specific proposed updates to the two exams and the research routinely and iteratively conducted in the cause of FCE and CAE validation as a whole clearly overlap. Nor is it straightforward to assign departmental ownership to many of the studies affecting the updates, since such studies are often collaborative across groups. Nevertheless, the qualitative and quantitative studies related to the two exams covered by the project were extensive and varied.

At the broad level of research for the updated FCE and CAE exams, we note in particular:

- specifications of VRIP (validity, reliability, impact and practicality) conditions for all papers of both exams according to a template following conventional interpretations of validity, but with insights from the Weir (2005) socio-cognitive framework for validating tests
- specifications for the updated examinations as mapped to the CEFR levels
- chair of paper overview reports for all ten FCE and CAE component papers.

At the fundamental level of exam *construct* research, Geranpayeh (2005), and Geranpayeh and Somers (2006) have conducted ground-breaking research. This seeks an inferential-statistical validation of the Cambridge ESOL model of language proficiency as deriving from a componential communicative language ability model, where each component, that is each of the five tests (Reading, Writing, Use of English, Listening and Speaking), assesses a different aspect of language proficiency.

In this issue of *Research Notes*, we are informed in greater detail of specific projects undertaken within the review of FCE and CAE examinations, including an overview of research undertaken on pages 31–4.

Information dissemination for the review

Numerous presentations have been arranged for stakeholder groups in major FCE and CAE exam constituencies during the latter period of the Review Project to ensure that key stakeholders were well informed of developments. These presentations, at key Cambridge ESOL centres or at conferences including IATEFL and BAAL, complement the information disseminated by the *Bulletins* on reviewing FCE and CAE, review updates in *Cambridge First*, the ESOL Website, teachers' seminars, the *Specifications and Sample Papers* documents (Cambridge ESOL 2007) and so on.

By late 2006 and into 2007, such audiences were hearing that the FCE/CAE Review was not 'a revision' involving radical change but was rather an 'updating' to ensure that the exams continued to meet customer requirements. Analyses of the crucial routine Candidate Information Sheets (CIS) were indicating, for example, that the candidature for CAE had changed since 1991.

Likely benefits from the FCE and CAE Review were now being summarised in the presentations and through other information dissemination means as follows:

- more straightforward progression from FCE to CPE, with FCE, CAE and CPE structured more similarly, 'thus encouraging candidates to progress from one level to the next: from CEFR B2 to CI to C2'
- reduced exam length, more appealing and accessible for test takers, perhaps enabling the whole of the exam to be taken in a single day (possibly with the exception of the Speaking test)
- additional results information to help candidates understand how they have performed, and help exam users in the interpretation of results.

A further purported benefit was 'an updated format to help teachers and students with exam preparation' a further reminder of Cambridge ESOL's historical concern for *exam washback.*

Face-to-face contact, through international presentations, seminars and invitational meetings, has been made with 1,244 stakeholders; there have been 14,000 website hits on the proposed FCE and CAE *Specifications*, with similar numbers of hard copies of the *Bulletins* and of the *Specifications and Sample Papers for examinations from December 2008* distributed. Standard seminar presentation packs have already been developed for use at appropriate public events. A promotional DVD with video clips, speaking packs for classroom use, FAQs on the website and further seminars for teachers, including teaching tips, are also provided. Feedback on the proposed FCE and CAE updates is considered to be very positive.

Updating FCE and CAE

Bulletin 5 (Cambridge ESOL 2006c) notes updates to the two exams which are of detail rather than radical, but which follow the suggested trends towards *exam family resemblance* and *improved time-effectiveness*. Some of the original areas of potential change have clearly been updated or deferred by the intensive consultation, research and decision-making processes of the project. Table 2 (based on *Bulletin 5*) summarises the changes to be made to FCE and CAE exams from December 2008 together with the rationales behind these changes.

It will be noted that the aim to reduce exam time has been pursued, as has that of standardising the format of the two exams. Other task changes and innovation are the result of both project-specific and routine research and of stakeholder feedback.

Specifications

The exam specifications summarised in *Bulletin 5* (Cambridge ESOL 2006c) are taken from the *FCE (or CAE) Specifications and Sample Papers for examinations from December 2008* (Cambridge ESOL 2006 d and e). These are characterised by their emphasis on validated exam and test *constructs*, as clear as possible definitions of target proficiency in terms of both ALTE and CEFR levels, and by an attempt to specify exam texts and tasks precisely. Such characteristics reflect the historical search for increasingly rigorous, valid and transparent exam specifications.

For both FCE and CAE exams the specifications include:

- content of Cambridge ESOL General English UMS examinations
- exam recognition, candidature, content and processing, marks and results and administrative information
- papers 1 to 5: General Description: (format, timing, number of parts and items; task and text types; lengths of texts; answer formats; scoring); and Structures and Tasks: by Parts: task type and focus, format, number of questions
- full sample papers with answer keys.

The content part of the specifications is crucial as it presents in a precise and transparent manner the theory, constructs and components of Cambridge ESOL Upper Main Suite exams. The presentation explains the 'skills and components' view of language proficiency accepted by Cambridge ESOL, the four skills of Reading, Writing, Listening and Speaking being themselves multi-dimensional involving the interaction of a language user's cognitive mental processing capacities with their language and content knowledge. All this, of course, in a purposive socio-communicative context. A fifth language component, 'Use of English' in test terms, focuses on the 'language knowledge structures or system(s) that underpin a user's communicative language ability', and includes knowledge of vocabulary, morphology, syntax, punctuation, and discourse structure.

Note the recurrence of our three themes which characterise the steady evolution of FCE and CAE: *exam levels* and *specifications* explicitly, *stakeholder relations* implicitly, as the target of the updated and intensified specifications documents.

Conclusion

This article has summarised, in its historical context, the FCE and CAE Review Project, a project that is still in progress and whose final outcomes will be exemplified in

Table 2: Changes and rationale for FCE and CAE updates

	Changes	Rationale
FCE		
Paper 1 Reading	• reduced from 1 hour 15 minutes (35 questions) to 1 hour (30 questions) • remove current Part 1 (matching headings or summary sentences to paragraphs of a text) • broaden scope of multiple- choice tasks on Part 2 text • use Part 3 gapped text task only in sentence form (not paragraphs).	• the new format is shorter, but has a broader test focus • item numbers will be fixed by part which simplifies the format • no alternative tasks will be available, which will make it more accessible for preparation.
Paper 2 Writing	• reduced from 1 hour 30 minutes to 1 hour 20 minutes • include production of emails • reduce word limit for Part 1 (response to long-input situation) • include review task in Part 2 • reduce set texts from five to two, with questions related to specific book.	• writing more briefly is a relevant real life skill, so the output required from Part 1 is reduced. The letter or email are compulsory because writing these is an important skill relevant to the candidature • the writing of email messages is added because they are used in modern life • a review task is added to Part 2 to increase the choices and broaden the candidates' experience • the set text questions will be related to a specific book • set books are reduced from five to two, as not all are well used.
Paper 3 Use of English	• reduced from 1 hour 15 minutes (65 questions) to 45 minutes (42 questions) • remove Part 4 (error correction).	• four of the five current tasks remain the same • the grammar that the error correction task tests is tested elsewhere in the paper • the skill of error correction can be tested through the writing paper.
Paper 4 Listening	• standardise format with other ESOL general English listening tests: i.e. one task type per section.	• a single task type in each section will improve comparability between versionsof the test and standardise the candidate experience. Candidates will know exactly which tasks to expect • the sentence completion task will be retained in Part 2 as this is more suitable to the level than note-taking • For Part 4, the multiple-choice task will be the only task retained; ensuring reliability between versions and over time.
Paper 5 Speaking	• add questions to visuals page in Parts 2 (long turns) and 3 (inter-candidate conversations) • add additional interlocutor prompts to Part 4 (discussion).	• adding questions to the visuals page will help candidates to recall the tasks. This will standardise their responses and improve the comparability of the tasks • Part 4 is enhanced with additional prompts for the interlocutor.
CAE		
Paper 1 Reading	• reduced from 1 hour 15 minutes (c. 45 questions) to 1 hour 15 minutes (34 questions) • remove Part 1 multiple-matching task • introduce themed texts (as in CPE Paper 1), wider sources, text range, task focuses.	• the range of text types, sources and task focuses in the paper is widened by the addition of themed texts in Part 1 • currently there are two matching tasks with some similarities, so Part 1's matching task is replaced by the themed texts.
Paper 2 Writing	• reduced from 2 hours to 1 hour 30 minutes • reduce Part 1 (article, report or proposal) input and output • add set text questions in Part 2.	• writing more briefly is a relevant real life skill, so the output required from Part 1 is reduced • Part 1 input is reduced to a number of words that falls between FCE and CPE requirements to reduce processing required • the nature of task output remains the same • trialling and research into shorter input and output for Part 1 shows that candidates can be successfully assessed with this length of words • reduction in input and output were found to be a positive development in research with examiners on the new tasks.
Paper 3 Use of English	• reduced from 1 hour 30 minutes (80 questions) to 1 hour (50 questions) • change name from English in Use to Use of English • remove register transfer, gapped text and error correction tasks • add gapped sentences and key word transformations.	• these changes will result in a more streamlined paper, which will be more similar to other Cambridge ESOL General English exams • key word transformations and gapped sentences are introduced: key word transformations appear in other levels and are efficient at testing grammar and vocabulary; gapped sentences proved successful in CPE, testing vocabulary productively • as register transfer and error correction can be tested through the Writing paper, both of these tasks are removed • research into the CAE construct shows that the gapped text task has a similar test focus to the gapped text in Reading; its removal does not reduce the range of what is being tested.
Paper 4 Listening	• reduced from 45 minutes (30–40 questions) to 40 minutes (30 questions) • introduce fixed format • include short extracts as in FCE and CPE • only one productive task • candidates to hear all texts twice.	• a fixed format enhances standardisation between versions • a standard format will appeal more to candidates • short extracts will be introduced in Part 1; these work well in FCE and CPE as they test a wide range of focuses and introduce a range of texts, interaction patterns and topics • the section which is once heard will be heard twice, following consultation and academic advice • the matching task will be retained as it discriminates well and tests gist listening effectively.
Paper 5 Speaking	• Part 1 (interlocutor conversation) two sections (instead of three) • remove section 2 of current Part 1 (candidates invited to 'ask each other something about...') • include questions on visuals page to support candidate (as in FCE).	• for Part 1, the new approach is more natural and will fit better with the other Speaking tests in the Cambridge ESOL Main Suite • as in FCE, Parts 2 and 3 will have questions provided on the visuals page to assist candidates in recalling the task • Part 4 is enhanced with prompts for the interlocutor.

the FCE and CAE exams of December 2008[1]. In the remainder of this issue, accounts appear of key research and other activities related to the exam updating process, covering whole exams (e.g. using factorial analyses modelling to explore the relationship of CAE papers to each other); concentrating on individual skills papers (e.g. updating or adding new tasks to FCE and CAE Reading, Writing, Listening, Use of English or Speaking papers) or updating another aspect of the exams such as the mark schemes for Main Suite and Business English Certificates.

References and further reading

Association of Language Testers in Europe (ALTE) (1998) *Handbook.*[2]

Bachman, L (1990) *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press.

Bachman, L, Davidson, F, Ryan, K, and Choi, I-C (1995) *An investigation into the comparability of two tests of English as a Foreign Language*, *The Cambridge TOEFL Comparability Study*, (Studies in Language Testing, volume 1), Cambridge: Cambridge University Press/UCLES.

Cambridge ESOL (2005a) *Reviewing FCE and CAE, Bulletin 1, May 2005*, Cambridge ESOL.

—(2005b) *Reviewing FCE and CAE, Bulletin 2, November 2005*, Cambridge ESOL.

—(2006a) *Reviewing FCE and CAE, Bulletin 3, February 2006*, Cambridge ESOL.

—(2006b) *Reviewing FCE and CAE, Bulletin 4, June 2006*, Cambridge ESOL.

—(2006c) *FCE and CAE – Exam specifications from December 2008, Bulletin 5, November 2006*, Cambridge ESOL, available from www.cambridgeesol.org/support/dloads/general/fcecae_review5.pdf

—(2007a) *Reviewing FCE and CAE, Bulletin 6, January 2007*, Cambridge ESOL, available from www.cambridgeesol.org/support/dloads/general/fcecae_review6.pdf

—(2007b) *Reviewing FCE and CAE, Bulletin 7, May 2007*, Cambridge ESOL, available from www.cambridgeesol.org/support/dloads/general/fcecae_review7.pdf

—(2007c) *Reviewing FCE and CAE, Bulletin 8, August 2007*, Cambridge ESOL, available from www.cambridgeesol.org/support/dloads/general/fcecae_review8.pdf

Chamberlain, A (2004) *Main Suite Modification (MSM) Market Research Report on the FCE and the CAE exams*, Cambridge ESOL internal report.

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

Davies, A (in press) *Assessing Academic English: Testing English proficiency 1950–1989 – the IELTS solution*, (Studies in Language Testing, volume 23), Cambridge: Cambridge ESOL/Cambridge University Press.

Hawkey, R (2004) *A Modular Approach to Testing English Language Skills*, *The Development of the Certificates in English Language Skills (CELS) Examinations,* (Studies in Language Testing, volume 16), Cambridge: Cambridge University Press/UCLES.

Huxley, A (1952) *The Devils of Loudun*, New York: Chatto & Windus/Harper & Brothers.

UCLES (1939) *Regulations*, UCLES.

Vidakovic, I (2007) *A summary of research activities associated with FCE and CAE modifications*, Cambridge ESOL internal report.

Weir, C and Milanovic, M (2003) (Eds) *Continuity and Innovation: Revising the Cambridge Proficiency Examination in English 1913–2002* (Studies in Language Testing, volume 15), Cambridge: UCLES/Cambridge University Press.

[1] For further information on the updated specifications, please visit www.CambridgeESOL.org/exams/fce.htm

[2] For information on current ALTE publications, visit www.ALTE.org

Using Structural Equation Modelling to facilitate the revision of high stakes testing: the case of CAE

ARDESHIR GERANPAYEH RESEARCH AND VALIDATION GROUP

Introduction

One of the essential considerations in the review of the CAE examination was to investigate what impact any changes in the format of the papers would have on the underlying constructs of the test. Geranpayeh (2005) argues that the operational definition of language proficiency in Cambridge Upper Main Suite (UMS) examinations is based on the notion that while there exists overall communicative language ability, such ability is divisible by skills and language elements. Since each skill can be developed to different degrees or at different rates, they can be separately recognised and measured. Hence there are four skills focused papers and a Use of English paper.

A careful examination of the content of each paper as illustrated in Weir & Milanovic 2003 (SiLT volume 15 on the revision of CPE) reveals that each Cambridge UMS paper assesses a different aspect of the overall communicative language ability at a particular level. That is, each paper provides a unique contribution to the building of a profile of communicative language ability that defines what a candidate can do at a level. The overall proficiency is built up from the combination of individual skills as measured by each paper.

The grading of UMS examinations reflects the underlying assumptions of Communicative Language Ability. Each paper is assessed and graded separately based on candidates' performance on each skill, the results of which are then added up to form a composite score. This aggregated score will be used to determine the adequate performance of pass at that level along the Cambridge ESOL Common Scale.

In short, the Cambridge operational definition of language proficiency falls within the eclectic approaches to language assessment. In this model, elements of communicative language ability, overall proficiency and its divisibility to language skills are married up to form a communicative model of language proficiency.

Empirical evidence to support the Cambridge UMS model

To examine whether the empirical evidence supported the assumptions made above, Geranpayeh (2005) analysed candidates' performance from a June 2004 live administration of CAE. Candidates' scores on all five papers were collected. The candidates' scores on each section of the papers were computed. For objective papers, the test sections as laid out in the paper were used as the basis of score computation, e.g. four sections for the Reading and the Listening papers and six sections for the Use of English Paper. The Writing paper was divided into the scores for the first and the second question while the speaking scores were divided into the five speaking assessment criteria. Hence 21 scores were collected for each student. There were altogether over 30,000 candidates in the analysis. However, for ease of analysis only scores from the students who did the listening version A were selected bringing down the total number of candidates in the sample to just over 11,000. The rest of the candidates were used for verification studies which were followed up later.

Using various Exploratory Factor Analysis techniques, plausible construct models for each paper were constructed. The viability of each model was tested by Structural Equation Modelling (SEM) techniques. A language proficiency measurement model was gradually constructed by adding one component at a time to the model. The first measurement model was a one factor Reading Comprehension model. Other objective papers such as the Use of English and Listening papers were gradually added to the equation and their plausibility was tested. Finally, the performance testing papers (Writing and Speaking) were added to the measurement models and plausibility of various models was tested using SEM techniques. The Best Fit indices came from a Correlated-Trait model as shown in Figure 1.

Figure 1 illustrates that the Correlated-Trait (CT) model is the best description of the CAE examination. That is, the Cambridge model of language proficiency is based on a componential aspect of communicative language ability whereby each component assesses a very different aspect of language proficiency. This also implies that apart from two sections, which we will shortly discuss, the convergent and discriminant validities – as illustrated by high/low correlations between test components – do not warrant the merging of any of the current papers as had been mooted at the outset of the review of CAE.

Implications for the review project

We tried to test measurement models that combined the Reading and the Use of English (UOE) papers in an attempt to simulate what was suggested for IELTS in 1992–5. Such models were consistently rejected in this study. It appeared that the overlap between these two papers was not higher than that of these papers and the Listening paper. In fact the overlap between UOE and Listening papers was slightly higher than that of UOE and Reading. The Use of English paper appears to test a unique aspect of proficiency which is different from that tested by the Reading paper. Looking at the correlations between various papers as illustrated in Figure 1, one may conclude that the grammatical ability tested by the UOE paper is very much a written feature. It has a very low correlation with the grammatical ability tested in the Speaking paper. We may say therefore that the UOE paper measures some aspect of written grammatical ability which is associated with Reading and Listening and to some extent with the questions in the Writing paper but not with the grammatical element assessed by the Speaking paper.

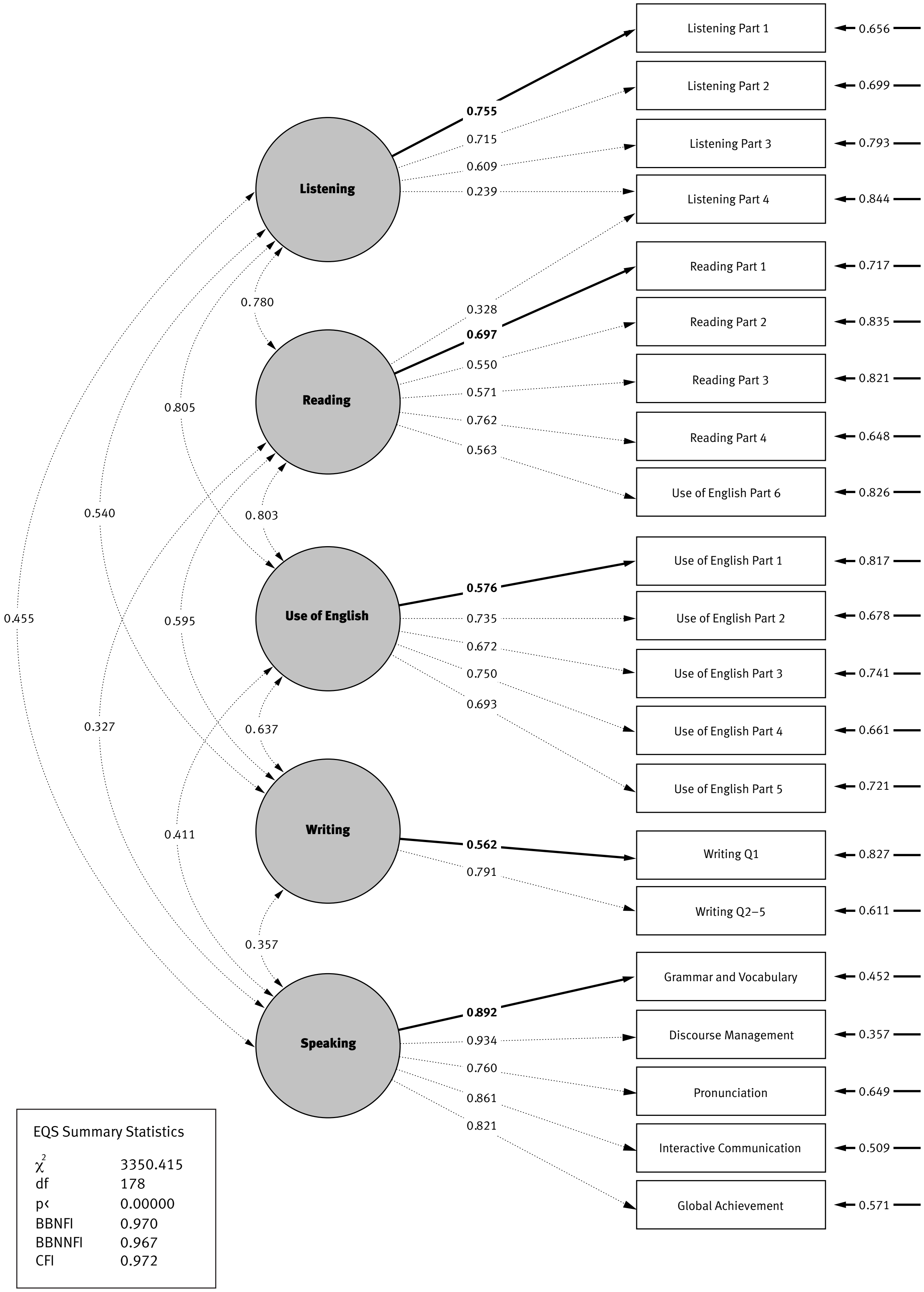
Having said the above, it is important to report that throughout this study we observed that Part 6 of the UOE paper consistently loaded very highly on the factor associated with the Reading sections. A content analysis of UOE Part 6 revealed that this section is testing coherence and cohesion, which are highly associated with discourse and hence Reading Comprehension. Although UOE Part 6 tests coherence and cohesion at a more micro level than that tested in the Reading paper, the study recommended the removal of this section from the UOE paper.

It was also observed from the initial study that Part 4 of the Listening paper (a multiple-matching type task) had a higher loading on the Reading factor than on the Listening factors. This task appeared to be an integrative task which measured both reading and listening. A content analysis of this section revealed that Part 4 had a very high load of reading factor embedded in its design. We further tested this hypothesis using data from versions B and C. The data from version B verified our hypothesis whereas the data from version C rejected it. The content analysis of Listening version C June 2004 showed that a different task (multiple-choice) was used in Part 4 which did not have the high reading load in its design. The current CAE test design specification allows the possibility of different listening tasks for the last part of the test.

In a separate study, Geranpayeh & Somers (2006) examined the performance of CAE candidates on a different administration of the test where three similar Part 4s were used in the Listening papers. Table 1 demonstrates that the Listening Section 4 in June 2005 did not load as highly on reading as it did in the previous study, shown by the higher factor loading (0.484) of this section on the Listening Factor rather than the Reading Factor. Further analysis of two other versions confirmed the above findings. That is, the different task types in Section 4 of the Listening paper primarily test listening comprehension although it is sometimes difficult to separate the influence of the reading factor on the multiple-matching integrative task.

Table 2 summarises the FIT statistics in support of the Correlated Model using six administrations of the CAE in June 2004 and 2005. The most meaningful row in Table 2 is the CFI. All the CFI values in the table are above 0.960 which are considered to be a very good fit for the measurement models tested in the studies.

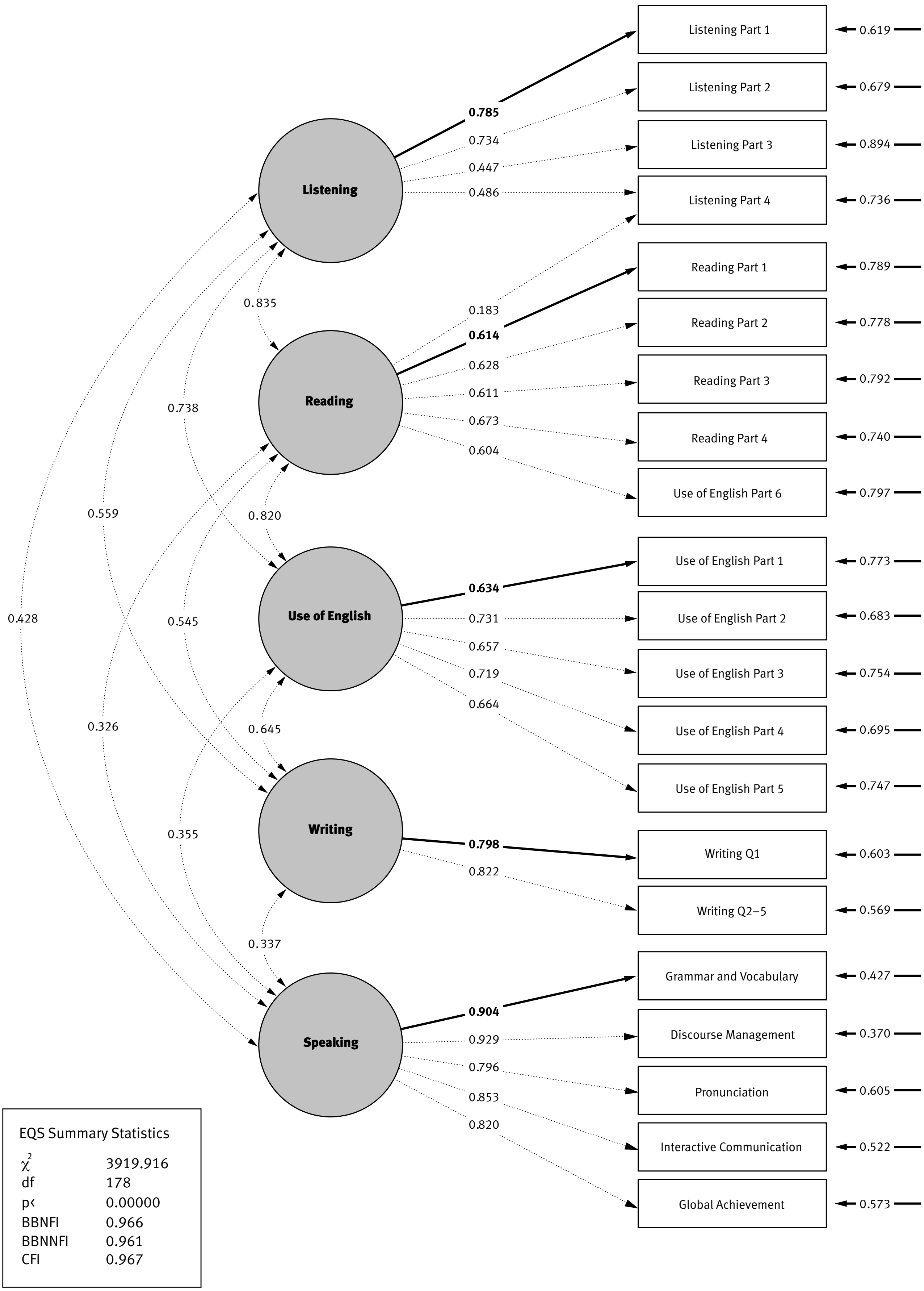
Figure 1: Correlated-trait CT model of June 2004 CAE exam (Listening Version A)



EQS Summary Statistics

χ^2	3350.415
df	178
p‹	0.00000
BBNFI	0.970
BBNNFI	0.967
CFI	0.972

Figure 2: Correlated-trait CT model of June 2005 CAE exam (Listening Version A)



EQS Summary Statistics

χ^2	3919.916
df	178
p‹	0.00000
BBNFI	0.966
BBNNFI	0.961
CFI	0.967

Table 1: Factor loadings (Direct Oblimin Solution) – June 2005 data using Listening Version A

Test Sections	Factor Loadings				
	Reading Factor	Speaking Factor	Use of English Factor	Writing Factor	Reading Factor
Reading Section 1	**0.679**	-0.004	0.105	-0.003	-0.101
Reading Section 2	**0.505**	0.001	-0.030	0.033	0.207
Reading Section 3	**0.564**	0.029	-0.004	0.067	0.056
Reading Section 4	**0.460**	0.039	0.113	-0.010	0.182
Writing Task 1	-0.002	-0.008	-0.003	**0.793**	-0.038
Writing Task 2	-0.032	0.006	0.010	**0.777**	0.004
Use of English Section 1	0.150	0.099	**0.274**	0.093	0.231
Use of English Section 2	0.220	0.057	**0.437**	0.092	0.065
Use of English Section 3	-0.018	-0.008	**0.676**	0.062	0.030
Use of English Section 4	0.168	0.016	**0.642**	0.041	-0.084
Use of English Section 5	0.024	0.006	**0.628**	0.044	0.059
Use of English Section 6	**0.390**	0.031	0.127	0.076	0.115
Listening Section 1	0.038	0.078	0.194	0.042	**0.560**
Listening Section 2	-0.066	0.068	0.298	0.029	**0.549**
Listening Section 3	0.118	-0.012	-0.076	0.052	**0.481**
Listening Section 4	0.239	0.074	-0.084	0.085	**0.484**
Speaking Grammar & Vocabulary	-0.012	**0.851**	0.023	0.012	0.009
Speaking Discourse Management	0.018	**0.886**	0.001	-0.018	-0.033
Speaking Pronunciation	-0.050	**0.804**	-0.060	0.027	0.041
Speaking Interactive Communication	0.016	**0.867**	-0.020	-0.036	-0.060
Speaking Global Achievement	0.003	**0.795**	0.027	0.013	0.015

Conclusion

In this article we argued that the Cambridge ESOL operational definition of language proficiency for UMS examinations is based on the notion that while there exists overall communicative language ability, such ability is divisible by skills and language elements, thus we have four skills papers and a Use of English Paper. We also argued that each UMS exam assumes a certain level of proficiency level and hence expects candidates to have overall language skills that level.

We reported that the CAE empirical data supports a Correlated-Trait model of language proficiency where each assessment component measures a very different aspect of communicative language ability. In other words, each component has a unique contribution to the assessment of the overall proficiency (pass/fail), i.e. they are all necessary for arriving at the composite score.

It was demonstrated that there is little information redundancy as assessed by each component that merits the merging of any two components. As a result of this, the future enhancement to the CAE exam should not change the five component structure of the examination. Having said that, the studies reported here recommended removing Task 6 from the UOE paper which is testing cohesion and coherence, a concept which is already being tested in the Reading paper.

We argued that the multiple-matching task in Part 4 of the Listening paper is an integrative task testing both Reading and Listening, more strongly the latter. To avoid comparability issues the study recommended fixing task types in different versions of the Listening paper and paying close attention to the number of words used in the distractors.

Finally it was argued that the UOE paper is testing grammatical ability associated with written discourse.

References and further reading

Geranpayeh, A (2005) *Building the construct model for the CAE examination*, Cambridge ESOL internal report.

Geranpayeh, A & Somers, A (2006) *Testing the construct model for the CAE examination*, Cambridge ESOL internal report.

Weir, C and Milanovic, M (2003) (Eds) *Continuity and Innovation: Revising the Cambridge Proficiency Examination in English 1913–2002,* (Studies in Language Testing, volume 15), Cambridge: UCLES/Cambridge University Press.

Table 2: Correlated Hypothesis Measurement Model of Language Proficiency FIT statistics

	June 04 A	June 04 B	June 04 C	June 05 A	June 05 B	June 05 C
No. Cands.	11506	9586	9565	11629	9336	10559
χ^2	3350.415	2518.16	2670.868	3919.916	3714.852	4543.166
DF	178	178	178	178	178	178
BBNFI	0.970	0.973	0.972	0.966	0.961	0.958
BBNNFI	0.967	0.971	0.969	0.961	0.956	0.953
CFI	0.972	0.975	0.974	0.967	0.963	0.960

χ^2 = Chi Square test

DF = Degree of Freedom

BBNFI = Bentler-Bonett Normal Fit Index

BBNNFI = Bentler-Bonett Nonnormal Fit Index

CFI = Comparative Fit Index

Introducing short themed texts into the CAE Reading paper

HELEN COWARD ASSESSMENT AND OPERATIONS GROUP

Introduction

This article outlines the process of consultation, discussion, trialling and review undertaken in updating the CAE Reading paper. The project conducted considered the CAE Reading paper as a whole, and formed part of the larger review of the FCE and CAE examinations. In this article, while some reference is made to how the review project has affected all parts of the Reading paper, the focus is the review of Part 1 of the paper, since it is this part which has seen the greatest development in the review process. To begin the article, a summary outlining the key issues arising from the initial survey and consultations is provided, followed by an account of how short themed texts came to be determined as appropriate tasks for inclusion in Part 1 of the CAE Reading Test Specification from December 2008.

Initial survey and consultations

In the initial stages of the review of the FCE and CAE papers, Cambridge ESOL conducted market research to obtain feedback from stakeholders on possible changes to both these levels of the Main Suite examinations. The content of the survey questionnaires revealed views on topics such as the length and content of the examinations. Two key issues dominated the feedback with regard to the CAE Reading paper:

- it was felt to be somewhat lengthy, particularly the length of the texts in Part 4 of the paper; relevant reading skills were felt to be testable with shorter texts
- having two multiple-matching tasks in the same paper was considered unnecessary.

In addition to the initial survey, the Chair of CAE Reading was commissioned to write an overview report of the paper, focusing on areas for consideration and on possible ways of updating the Reading paper. Following this, a meeting to discuss key issues arising from the report and the initial survey was held internally, attended by external consultants and internal staff. At this meeting, it was reported that all tasks in the CAE Reading paper were working well; however, it was thought that having one multiple-matching task rather than two would impact positively on the face validity of the paper. Furthermore, it was agreed that the appropriate level of challenge could be maintained in a shorter paper, pending satisfactory trialling. While consultants believed that a reduction in the number of items to roughly 35 would bring CAE Reading more in line with the progression from FCE to CPE, it was not felt that the number of tasks found in the CAE Reading paper should be reduced, as it was important to maintain the reliability of the paper and for candidates and teachers to see a clear development from FCE through to CPE in terms of length of texts and demands of the type of tasks required. Table 1 shows the current CAE Reading test format. Data collected on candidature had revealed that a significant proportion of FCE candidates go on to take CAE and a number of these then progress to CPE. Thus, one of the aims of the review of the paper was to streamline it so that it reflected a more straightforward progression from FCE through to CPE.

Table 1: CAE Reading test specification 1999 – June 2008

Part	Task Type	No. of Questions	Task Focus
1	Multiple-matching	12–18	Specific information
2	Gapped text	6 or 7	Text structure
3	Multiple-choice (long text)	5–7	Detail, gist, opinion or attitude
4	Multiple-matching	12–22	Specific information

Test length: approximately 45 items. Test time: 1 hr 15 m

Explorations of other possible Part 1 tasks

The next stage of the development of the review of the CAE Reading paper concerned the exploration of possible tasks which would be suitable to take the place of the existing multiple-matching task in Part 1, while maintaining the appropriate level of challenge, and increasing, if possible, the range of test focuses. Two initial suggestions were given careful consideration by external consultants and internal staff. The merits and drawbacks of each were taken into account, the details of which are outlined below.

Multiple-matching with headings

One possibility under consideration was to replace the current multiple-matching task found in Part 1 with a different type of multiple-matching task; one in which candidates would be required to match headings to sections of text, as in Part 1 of the current FCE Reading paper. This would not only demonstrate a closer link with the FCE paper, but it would also have reading for gist as its testing focus. It was noted that there were only a few items in the current paper that had this type of expeditious reading as the testing focus. However, an analysis of how this type of task was performing at FCE level revealed definite drawbacks. It was found that, at FCE, multiple-matching tasks with headings often fell below the target ability level, and so led to item writing issues. Taking this into account, and placing emphasis on the importance of maintaining the level of the CAE Reading paper, it was decided not to proceed with this proposal.

Lexical cloze

The proposal to include short lexical cloze passages, a task in which candidates complete a short gapped text by selecting the correct word or phrase from a set of four options, similar to the CPE Reading Part 1 model arose from the acknowledged desirability of having an increased number of items with a focus on testing lexis in CAE Reading. Apart from the fact that modelling such a task on the one found in Part 1 of the CPE Reading paper would bring CAE Reading more in line with the development to CPE, it was well documented that this type of task was a successful discriminator in terms of level. A few lexical cloze passages were commissioned for the trial tests and were found to perform successfully. However, when considering the review of CAE as a whole, it was decided that the testing of lexis would be adequately covered in the Use of English paper and so, it was felt that the CAE Reading test would be better served by introducing a task which would have scope for testing a wider range of focuses than could be achieved in a lexical cloze.

Following the consideration of the above two tasks, attention turned to the following task which appeared suitable in terms of meeting the testing requirements of the updated CAE.

Short themed texts

Short themed texts is a type of task used in Part 2 of the CPE Reading paper and one which evidence suggested has been performing successfully at this level. The proposal to introduce this type of task to the CAE Reading paper was based on the following merits:

- The range of Reading test focuses would be widened.
- It would enable a broader range of genres and registers to be used in the paper. Attention was drawn to this particular merit in a report written by the Chair of CPE Reading in which it was suggested that this broader range would allow for a good washback effect on preparation materials and teaching.
- It would bring CAE more in line with CPE, encouraging clearer development from CAE to CPE.

Also stated in this report are the further two positive elements regarding short themed texts:

- Candidates are given a fresh start with each new text
- The use of themed texts supports the topic-based nature of many course books and of classroom practice.

All in all, the introduction of short themed texts would, as had been the case at CPE level, allow a 'wider sampling of text types, topics and reading strategies' (Ashton in Weir and Milanovic 2003:133).

The texts would come from a variety of genres, including journalistic articles from newspapers and magazines, journals, books (fiction and non-fiction), ephemera such as promotional and information materials, and the testing focuses would include detail, opinion, tone, purpose, main idea, implication, attitude, and text organisation features (exemplification, comparison, reference). In order to differentiate the task from that found at CPE, it was decided that there would be three, rather than four themed texts. Each text would be followed by two four-option multiple-choice questions on the text, totalling six items for the three texts. The content of the texts chosen for CAE would also be slightly different in nature from those at CPE in that they would be less abstract, literary or philosophical than those which may be used in the CPE Reading paper.

It appears from the above that a strong case could be made for the use of short themed texts as a new Part 1 in the CAE Reading paper. However, in order to establish whether this type of task would be appropriate for use at CAE, a number of trial tests and further consultations were required.

Trial tests and trial test reviews

Between June 2005 and April 2006, four CAE Reading trial tests were developed and trialled. Tasks developed for use in these trial tests included a number of short themed texts, which were specially commissioned, written by experienced CAE Item Writers.

The trial tests were completed by prospective CAE candidates from a wide variety of countries: Argentina, UK, Italy, Libya, Switzerland, Poland, Brazil, Germany, Malaysia, New Zealand, Ireland and Austria. Statistical data, as well as qualitative feedback from the centres and participants involved in the trial tests, was collected.

Trial review meetings attended by internal staff and external consultants were held following the trial tests. The statistical data gathered from each test, and the feedback given by those who had participated were carefully scrutinised at these meetings and reports written summarising the findings. These reports show that the tasks that were trialled had performed with a high degree of success. Tasks comprising short themed texts were found to be within the appropriate range of difficulty for CAE and had discriminated well. Moreover, it was agreed that there had been a good range of question focus and type of text in these tasks in each trial test.

Centre feedback indicated that short themed texts were appropriate for use at CAE level with comments on the positive features of the task type, including the observation that these themed texts cover a wide range of lexical and semantic fields, and that candidates who find a particular text challenging would have more chances of success as there is a greater variety of texts available.

Number of items and length of the CAE Reading paper

The results of trial tests also confirmed the most suitable number of items and length of texts in order to maintain the reliability of the Reading paper. The length of each part would be within the range of 550–850 words and the number of items on the paper would be fixed at 34, a reduction from the number in the current paper, which contains approximately 45 items (see Table 2 for updated test format). This responds to the feedback on length of CAE Reading, while at the same time giving due attention to maintaining the appropriate level of challenge in the type of texts selected and in the items written, so that reading will not be seen to have a diminished role.

Research into the weighting of questions

As a further part of the review process, research in the form of observation was also conducted to establish the appropriate weighting for each part of the updated paper. This research confirmed the proposed double weighting of questions in Parts 1, 2 and 3 of the updated Reading paper, each question receiving two marks, and the single weighting of Part 4, i.e. that each question in this part should receive one mark. The evidence was that candidates spend longer on Parts 1, 2 and 3 as they read in more depth than in Part 4.

Table 2: Updated CAE Reading test specification December 2008 onwards

Part	Task Type	No. of Questions	Task Focus
1	Multiple-choice 3 short themed texts	6	Detail, opinion, tone, purpose, main idea, implication, attitude, text organisation features (exemplification, comparison, reference).
2	Gapped text	6	Text structure, cohesion and coherence
3	Multiple-choice (long text)	7	Detail, opinion, tone, purpose, main idea, implication, attitude, text organisation features (exemplification, comparison, reference).
4	Multiple-matching	15	Specific information, detail, opinion and attitude

Test length: 34 items. Test time: 1 hr 15 m

Conclusion

The results of the trial tests strongly supported the introduction of short themed texts as a task in Part 1 of the updated CAE Reading paper from December 2008. It was clear from the feedback collected that most centres involved in the review process were in favour of their introduction, expecting them to impact positively on the Reading paper. It would indeed allow the range of testing focuses and variety of genres available for use in the paper to expand, thus bringing more scope into the assessment of reading ability at CAE level. Any concerns raised over possible problematic features of having short themed texts at CAE were addressed at trial review meetings, with the following recommendations:

1. It is important to stress that the theme connecting the texts should be broad.
2. The use of fiction would be introduced into the paper in general, including in this part. It should, however, be treated with caution, and alternative forms of fiction to literary texts, such as detective fiction and fiction found in magazine supplements are advised. It is also important to provide enough context for candidates to understand what is happening in the texts. A clear title should be given with fiction texts (as with all texts) in order to help the candidate establish the context of the text.

Results from all the research conducted during the development of the updated CAE Reading paper have shown that using short themed texts as a task at CAE level would fit into the proposed length of the updated paper of approximately 3,000 words. Timed trials were set up, observed, and showed that in view of the minimal change in word length of the whole Reading paper, and the fact that despite the reduction in the number of items, different types of task demand different lengths of time for their completion, the time limit for the whole paper would remain at 1 hour and 15 minutes.

Reference

Ashton, M (2003) The change process at the paper level. Paper 1, Reading, in Weir, C and Milanovic, M (Eds) *Continuity and Innovation: Revising the Cambridge Proficiency Examination in English 1913–2002,* (Studies in Language Testing, volume 15), Cambridge: UCLES/Cambridge University Press, 121–174.

Establishing the impact of reduced input and output length in FCE and CAE Writing

MARGARET COOZE ASSESSMENT AND OPERATIONS GROUP
STUART SHAW RESEARCH AND VALIDATION GROUP

Introduction

As an outcome to the initial consultation process, one of the aims of the FCE and CAE Review Project was to investigate the possibility of shortening the time for some of the papers, including Writing, in order to reduce the overall time for each examination.

An issue crucial to the assessment of writing is the amount of time test takers are assigned to complete a task or group of tasks (Weigle 2002:101). Test time necessarily includes time given over to processing any textual input. In addition to the amount of text the test taker is required to deal with is the actual number of words the test taker is expected to produce. Input text length may have an effect on performance in terms of the underlying cognitive processing undertaken by the candidate. In general, therefore, the longer the input text candidates are presented with, the greater the language knowledge that might be required to process it (Weir 2005:69).

This article reports on a series of validation trials conducted on FCE and CAE Writing Part 1 tasks. The trials involved a group of examiners (identified as representative of the 'universe' of examiners for the Cambridge ESOL FCE/CAE Writing tests) rating a sample of performances which were prompted by updated question types.

Research questions

A principal aim of the trials was to confirm both the reliability and validity of the FCE and CAE Writing tasks and to ensure that the writing assessment process could continue to be robust and consistent. The trial attempted, therefore, to facilitate an understanding of how the updated tasks function by addressing the following questions:

- Do the updated FCE and CAE tasks (with reduced input and reduced expected output) discriminate well amongst candidates?
- Can the reduction of written text expected of FCE/CAE candidates be justified?
- What impact does reduced input and output text have on the validity/reliability of the FCE/CAE Writing paper?
- To what extent does a shorter output allow examiners the opportunity to reliably assess candidate performance at the FCE/CAE level?
- Can FCE/CAE candidates complete the set task within the word length range at their level?
- Does the reduced FCE/CAE output allow candidates to demonstrate a range of structural and lexical ability?

Methodology

In outline, the procedure was to introduce groups of trial examiners to the updated tasks and then do multiple marking of sets of scripts. Multiple marking by a reasonably large number of examiners using the same scripts provides a large number of inter-rater correlations, which are a focus of interest.

It was hoped that the combined use of quantitative methodologies (application of general and updated task-specific criteria and scales to sample language performance) and qualitative methodologies (observations from 'expert' participants) would inform any refinements to the proposed Writing tasks.

The qualitative dimension of the trial comprised individual verbal protocols collected and recorded during actual marking. Personal examiner observations, insights and concerns were also captured throughout the marking event to supplement the verbal protocols. In addition, a plenary session was undertaken: this took the form of a semi-structured focused discussion group designed to enable the examiners to share their views with each other and with the Cambridge ESOL team. The open discussion nature of focus groups means that participants discuss their views in an atmosphere conducive to collaborative interaction.

The empirical dimension of the trial involved the collection and subsequent analysis of the ratings of trial participants. Quantitative methodologies included correlational analyses; computation of examiner inter-rater reliabilities; and Multi-Faceted Rasch Analyses.

Scripts were selected for the two trials to include a range of variables: candidates from a range of L1 backgrounds and ages; candidates studying in the UK and in monolingual situations; and, candidates following preparation courses (in the case of FCE) and those following general language courses.

Previous validation trials have required a minimum of five raters for re-rating purposes (Shaw 2002, Shaw and Galaczi 2005). In this trial, six examiners were used for the FCE trial and five examiners were used for the CAE trial. All the examiners were independent and variously experienced Cambridge ESOL examiners and included Team Leaders (TLs) with wide experience of FCE/CAE marking and of managing examiners during live marking sessions as well as Assistant Examiners (AEs) with differing amounts of marking experience on their respective papers.

FCE trial materials and participants

Feedback from the consultation process indicated that there was a strong desire to include emails as an alternative to the compulsory letter in Part 1 of the FCE Writing paper, as they were seen to be relevant to the candidature. Therefore, it was decided to update a letter task which had previously been used in a live administration and had performed extremely well. The updated task took the form of an email with four content points (a reduction from five in the current model as a result of the amended output word length required). The updated task contained one input text and required one output text. The total number of words for the rubric and input text was 184; the candidates were required to write their responses in the range of 120–150 words.

In total, eight participants took part in the trial: six examiners (two TLs and four AEs) and two Cambridge ESOL staff (a Subject Officer and a Validation Officer). The six examiners each rated a total of 30 common writing performances (i.e. 30 Part 1 responses to the updated task). The Part 1 responses were selected by the FCE Writing Subject Officer to reflect the variables noted earlier. Script responses constituted material from various time trials and included candidates following FCE preparation courses both in the UK and overseas, as well as candidates studying on General English courses in the UK. The cohort included children (from the age of 14) and adults. This ensured that the study replicated the candidature of FCE as a whole for live administrations.

CAE trial materials and participants

Part 1 of CAE has tended to require more text processing by candidates than FCE and CPE. In light of the information gathered about the change in candidature for CAE, Cambridge ESOL took the opportunity of rationalising this within the FCE and CAE Review Project.

As the changes to CAE Part 1 were more significant than to FCE, two Part 1 tasks were used for the CAE trial to explore the more significant reduction in input. Both tasks were similar in that they were 'proposals', containing one input text and requiring one output text; both contained inputs (rubric and text) of a very similar length (142 and 149 words).

The CAE trial involved seven participants: five examiners (two TLs and three AEs) and a Cambridge ESOL Subject Officer and Validation Officer. The five examiners each rated a total of 40 common writing performances (20 × task A, 20 × task B), once again selected by the CAE Writing Subject Officer to reflect the variables noted earlier (age, L1 etc). The trials took place in centres worldwide during October 2005 and February/March 2006.

Trial findings

Results of the quantitative analyses provide evidence of the validity and reliability of the updated writing tasks: specific conclusions gleaned from the trials can be related to the various statistical methods employed.

Although there were differences in overall severity between examiners, descriptive statistics and ANOVA indicated that the examiners were homogeneous in their marks. Whilst in terms of absolute scores examiners demonstrated some disparity in rating, any differences were marginal and for all practical purposes the examiner group can be thought of as being equally severe. Mean examiner intercorrelations and statistical significance tests indicated that the strength of correlation was such that there was evidence of good relationships between examiners. Inter-rater reliabilities were encouraging: of the order of 0.7. Multi-Faceted Rasch Analyses (FACETS) revealed that all examiners fell within the limits of acceptable model fit and that whilst examiners were not equally severe in their assessments any differences in severity (between the most and least severe) were small. A high degree of examiner consistency was also manifest in the data. Additionally, FACETS indicated that all examiners were operating within an acceptable range of consistency of performance. FACETS indicated only two ‘problematic’ ratings across the examiner group suggesting that individual examiners were ‘misfitting’ the Rasch model on only two occasions. FACETS revealed a wide range in performance across the script set.

Supplementary findings from complementary qualitative studies point to the practicality and positive impact of the updated tasks from the examiners’ perspective: examiners were enthusiastic about the trial claiming it to be a very positive and worthwhile experience. Verbal reports together with issues raised during the focused discussion revealed examiners were generally very favourably disposed to the updated FCE/CAE Writing Part 1 tasks considering them to be a positive development particularly in terms of perceived enhanced reliability of assessment.

Issues raised by examiners highlighted various aspects relating to the nature and content of the updated tasks. Reduced input and output text was welcomed, affording greater clarity to both candidates and examiners. Updated instructions were also perceived to provide enhanced lucidity and fairness to all candidates. Shorter input text appears to facilitate fewer opportunities for candidate ‘lifting’. In addition, it was believed that a reduction in the quantity of input text permits test takers and examiners to focus more on language: the production of language (candidate perspective) and the assessment of language (examiner perspective).

The issue of whether the sample of written performance produced by the test taker in response to the updated tasks is sufficient, i.e. of length generally accepted as a minimum, was prevalent in the minds of examiners. From a reliability standpoint, the updated tasks appear to engender a sample of writing appropriate for assessment purposes. Most examiners believe that scoring reliability is not compromised, i.e. candidates do not appear to be adversely affected, in terms of their scores, by writing less. However, examiners may well require time to familiarise themselves with a reduction in actual written output as the updated stipulations are currently at variance with the output they are accustomed to.

FCE findings

The use of email as a task type for the FCE writing paper was perceived as a very positive addition, and the task used in the trial was seen to be able to effectively discriminate between weak and strong candidates.

CAE findings

A general perception appeared to be that reliability of assessment was enhanced since it is now easier for examiners to retain task features in their minds during marking thus obviating the need to continually revisit input task material throughout the marking episode.

Part 2 questions

Whilst Part 2 questions were not subject to investigation in this trial, the task type range has nevertheless been updated. In Part 2 of the FCE Writing paper, ‘review’ has been added as it is felt to be relevant to the candidature and appropriate to the type of language that FCE level candidates can produce. This has proved popular in trialling and candidates have performed well on the tasks set. The word length for questions in this part remains unchanged at 120–180 words. In the CAE Writing paper, tasks which are more appropriate in a work context, e.g. memos, have been removed to reflect the change in candidature.

Feedback from the consultations showed that while only a small proportion of candidates opt to do a set text question in Part 2 of FCE and CPE, their retention at FCE level and their introduction at CAE level was very desirable. The use of set texts in general English classes provides a focus for language learning, and the selection of texts which have also been turned into films adds an extra dimension to their use.

FCE and CAE Writing specifications

Changes to the existing test specifications for FCE and CAE, together with the rationale underlying such changes, are given in Tables 1 and 2. The tables also indicate possible implications for Validity, Reliability, Impact and Practicality (VRIP).

Table 1: Current and revised FCE Writing specifications

FCE Writing: Current Specification

Timing/ Length	Task Types and Format	Test Focus
1 hr 30 m 2 parts	Part 1 – Compulsory transactional letter (120–180 words) Part 2 – Optional article, non-transactional letter, report, discursive composition etc. (120–180 words)	Part 1 – focus on expressing opinions, justifying, persuading, comparing, recommending, suggesting, advising, apologising, describing and explaining. Part 2 – varying focuses according to task, including: expressing opinions, justifying, comparing, recommending, advising, describing and explaining.

FCE Writing: Specification from December 2008

Timing/ Length	Task Types and Format	Test Focus
1 hr 20 m 2 parts	Part 1 – compulsory letter or email (120–150 words) Part 2 – Q2–Q4 choice of one task from the following: article, letter, report, essay, review, story. Q5 task based on set texts-task types from: article, essay, report, review, letter; two set books to be used, with a question on each text (120–180 words)	Part 1 – focus on expressing opinions, justifying, persuading, comparing, recommending, suggesting, advising, apologising, describing and explaining. Part 2 – varying focuses according to task, including: expressing opinions, justifying, comparing, recommending, advising, describing and explaining.

FCE Rationale

Change	Rationale	VRIP	
A reduction in suggested word length for Writing Part 1. 1 hr 30 m – 1 hr 20 m	Part 1 output is reduced by 30 words – brief writing is considered a relevant skill at this level. The letter is retained as a task type for the compulsory part, as it is most relevant to the candidature, and writing messages by email is added.	V R I P	– – + +

Key: + positive implications – remains the same

Table 2: Current and revised CAE Writing specifications

CAE Writing: Current Specification

Timing/	Task Types and Format	Test Focus
2 hrs 2 parts	Part 1 – Compulsory article, notice, formal/informal letter, directions, instructions etc. (250 words) Part 2 – As for Part 1, but optional (250 words)	Parts 1 & 2 – applying information contained in the input, selecting and summarising input, comparing information.

CAE Writing: Specification from December 2008

Timing/	Task Types	Test Focus
1 hr 30 m 2 parts	Part 1 – Compulsory task from: article, report, proposal, letter (180–220 words) Part 2 – Q2–4 choice of one task from the following: article, letter, report, proposal, review, competition entry, contribution to a longer piece, essay, information sheet – Q5 task based on set texts – task types from: essay, review, article, report; two set books to be used, with a question on each text (220–260 words)	Part 1 – focus on evaluating, expressing opinions, hypothesising, persuading Part 2 – varying focuses according to task: including giving opinions, persuading, justifying, giving advice, comparing.

CAE Rationale

Change	Rationale	VRIP	
Part 1 input and output is reduced and Part 2 retains its current format. 2 hrs/ 2 tasks– 1 hr 30 m/ 2 tasks	The input in Part 1 is reduced to a number of words that falls between FCE and CPE, as currently it requires considerable processing. The nature of the task in terms of output would remain the same. The reduction in time is based upon the reduced input and output, as well as observations during live examinations.	V R I P	– – + +

Key: + positive implications – remains the same

Conclusion

The Writing papers for FCE and CAE have been updated to reflect feedback from stakeholders, consultations with external specialists in the field of writing and Senior Examiners as well as the outcomes of research projects.

It is clear from the trial findings that the impact of reduced input and output length in Part 1 FCE and CAE Writing tests is generally very positive although examiner feedback would suggest that a reduction in text may have implications for task design. Both the reliability and validity of the FCE and CAE Writing Part 1 tasks were upheld by their updating reported here, ensuring that the assessment of writing at FCE and CAE will continue to be robust and consistent. The trial showed that the updated FCE and CAE tasks discriminated well amongst candidates and a shorter output allowed candidates to demonstrate a range of structural and lexical ability which examiners could reliably assess at the appropriate level. Moreover, the FCE/CAE candidates who took part in the trial were able to complete the set task within the word length range at their level, indicating that the reduction of written text expected of FCE/CAE candidates was justified.

This research confirmed that the updated FCE tasks can be constructed to:

- be sufficiently challenging for more able candidates
- comprise clearly discernible content points
- be assessed against clearly worded task-specific mark schemes.

Similarly, the updated CAE tasks can be constructed to:

- be as 'open' as possible
- be sufficiently challenging for more able candidates
- employ input language more conducive to paraphrasing
- attempt to remove any element of choice presented to candidates
- offer less support to candidates (by reducing the amount of input).

In general, feedback on the proposals indicates that test takers and test users are generally very satisfied with the format and content of the updated Writing papers for FCE and CAE.

References and further reading

Shaw, S D (2002) *The Effect of Standardisation Training on Rater Judgement and Inter Rater Reliability for the Revised CPE Writing Paper 2*, UCLES internal report.

Shaw, S D and Galaczi, E D (2005) Skills for Life Writing Mark scheme Trial: validating the rating scale for Entry Levels 1, 2 and 3, *Research Notes* 20, 8–12.

Weigle, S C (2002) *Assessing Writing*, Cambridge: Cambridge University Press.

Weir, C J (2005) *Language Testing and Validation: An Evidence-Based Approach*, Basingstoke: Palgrave Macmillan.

Reviewing the CAE Listening test

STEVE MURRAY ASSESSMENT AND OPERATIONS GROUP

Introduction

In this article, we offer a description of the process of consultation, discussion, trialling and research we undertook in reviewing the CAE Listening test, and also an account of how we reached the conclusions which led to the finalising of the updated test specification which will be implemented from December 2008. The review we conducted mirrors the Cambridge ESOL exam review cycle as described in the introductory article for the FCE and CAE Review Project. In terms of the scope of the project with reference to the CAE Listening test, we reviewed all aspects of the exam specification and investigated and trialled updated tasks on a total sample of more than 700 prospective CAE candidates. To begin this article, in order to provide the reader with an overview of the CAE Listening test, we present the existing specification for the test (Table 1); the new specification can be found at the end of this article in Table 3.

Table 1: CAE Listening test specification 1999–June 2008

Part	Task Type	No. of Questions	Task Focus
1	Sentence completion or note completion task Heard twice	8–10	Following the main points and retrieving specific information from the text.
2	Sentence completion or note completion task Heard once	8–10	Following the main points and retrieving specific information from the text.
3	Multiple-choice long-text or sentence completion task Heard twice	6–10	Understanding specific information, gist and attitude in the text.
4	Multiple-matching (2 parallel tasks) or 3-option multiple-choice task Heard twice	10	Identifying speakers and topics, interpreting context, recognising attitude and function, gist and main points.

Test length: 32–40 items.
Test time: approximately 45 mins (35 min listening, 10 min transfer time)

Initial survey

In the market research conducted by Cambridge ESOL, with regard to the CAE Listening test there was evidence from those who responded that:

- a shorter test would be welcomed if it could be achieved without impacting on the historical comparability and validity of the test
- all parts of the paper should be twice-heard
- the multiple-matching task was sometimes perceived to be a challenging task type.

These key points, identified by the research done at this initial survey stage, informed the efforts we went on to make on consulting further and trialling new and updated tasks for possible inclusion in the updated CAE Listening test.

Consultations

Also at an early stage, a report focusing on areas for consideration and possible modification was commissioned from the Chair of the CAE Listening paper. This report became a key document which informed our review process at all subsequent stages. Following the submission of this report, there were further meetings with consultants and internal staff in order to discuss the principles or aims which should underlie any updates, and to produce outline specifications for any proposed new or updated tasks. In Table 2 we present some of these general aims which, along with their anticipated outcomes, were agreed.

In the event, these broad principles for revising the test, along with their anticipated outcomes, guided our reflections during the cycle of the review process, as well as informing further consultation and invitational meetings with a range of stakeholders, including publishers.

Table 2: Aims and outcomes of the review of the CAE Listening test

Aims	Anticipated outcomes
• to produce a fixed format test rather than a test with alternative task types and a variable number of questions	• improve the accessibility of the test experience for the candidates • enhance, as far as possible, the continuing parity of versions across administrations • standardise the assessment experience for candidates, who will know which tasks to expect in each test.
• to broaden the range of testing focuses, tasks and text types	• widen the range of listening skills covered in the test construct • enhance the construct validity of the test (Buck 2001:153) • enhance the possibility of a positive washback effect on the learning experience for candidates.
• to shorten the overall time of the test without impacting on the reliability of assessment	• improve the appeal of the test for test takers and stakeholders.
• to update the structure of the test to help teachers who teach FCE, CAE and CPE	• offer a more straightforward progression from FCE to CPE by structuring the exams more similarly, thus encouraging candidates to progress from one level to the next.

Drafting a specification

Bearing the above principles in mind, in considering a draft specification for the paper we decided to focus on the following areas: reviewing alternative task types; updating existing task types and introducing a new task.

Reviewing alternative task types

Firstly, we reviewed the existing range of alternative task types, during which process, as mentioned, the Chair's initial report and continuing input were invaluable in informing our decisions. In terms of the existing alternative note-taking task, although we could observe that over its lifetime this had performed to our statistical criteria for testing listening at the level, we felt this task is perhaps better suited to testing at lower levels, just as had been the observation on the task in the context of the FCE review process. So, investigation into the retention of this task type on the CAE Listening test was not pursued.

Sentence-completion task

The existing test specification offers two incarnations of the sentence-completion task: once-heard and twice-heard variants. The sentence-completion task presents candidates with a set of statements on a listening text, from which the key information has been removed. The task requires the completion of gapped sentences with a key word or short phrase distinguished during listening. The words or short phrases recorded on the question paper are then transferred by the candidate, along with the other answers, to a separate mark sheet at the end of the test. In the Chair's initial report, the once-heard incarnation of this task had been flagged as a possible candidate for removal, a suggestion supported by feedback received on this task from candidates and teachers, which tended to focus on the perceived difficulty of hearing a listening text once and so having only one chance to listen and catch the answer. Additionally, it seemed that the concept of the once-heard task could become less relevant in the CAE context, given changes in technology which mean that people can generally listen to online materials, e.g. radio programmes, as often as they wish. In the end, our consensus was that the once-heard variant would not be retained in the updated CAE Listening test.

Despite this, we felt the sentence-completion task in its twice-heard format was a productive task which could continue to meet the desired criteria for the level, in terms of testing to the level and offering a skills focus on detailed listening. In terms of statistical evidence, we observed from live examination and pretesting data that the task consistently achieved the target difficulty for the advanced level, while also appearing to perform strongly as an instrument of assessment by discriminating well between more and less able candidates. However, in terms of the overall balance of the test, we decided that the sentence-completion task should be used in one part only, in order to allow for a range of task, text and testing focuses across a fixed format paper.

Multiple-choice and multiple-matching tasks

Turning to Part 3, we considered the existing long-text multiple-choice alternative for Part 3 was a task which it was desirable to retain. This task presents candidates with a set of options which describe different interpretations of the force or meaning of a discussion, or a speaker's utterances, from which candidates must select the option they feel best expresses what they hear. We felt this task could continue to provide an appropriate testing vehicle for focusing on the understanding of attitude and opinion at the CAE level, and that the balance of task, text and testing focuses on the paper would be supported by its inclusion. However, we did, as described below, investigate the possibility that it might do so in a slightly updated format.

The multiple-matching (MM) task in Part 4 had drawn some negative comments from both students and teachers regarding its format (two parallel tasks intended to be done simultaneously), despite a consistently robust statistical performance in testing listening at the advanced level. After considering this feedback carefully in reviewing the alternative tasks for Part 4, we decided that the value of retaining the MM task outweighed that of keeping the existing alternative 3-option multiple-choice task. In support of this decision, we felt that the testing focuses which the 3-option multiple-choice alternative task offered could be adequately covered elsewhere in the Listening test; we considered that the MM task would enable the retention of the underlying construct, the range of testing focuses, task and text types across the Listening test as a whole. In addition, the research into construct (reported earlier in this issue) suggested that only one task should be retained in this part of the test. However, we did feel it appropriate to review the format of the MM task, and so we went on to investigate thoroughly whether it might benefit from modification.

Updating the existing task types

With the long-text multiple-choice task, one proposal was that we should investigate whether 3- or 4-option multiple-choice questions (MCQs) would be most appropriate. We felt it was unnecessary to modify the text type or testing focuses of the task itself, given the strengths of the task as described above, so we decided the listening input would continue to be a long text, either a discussion or interview involving two or three speakers; and the items would continue to focus on the candidate's ability to comprehend the attitude and opinions of the speakers. However, we did consider it worth investigating whether we could reduce the processing load of the questions by reducing the number of options from four to three. Initially, we anticipated positive observations might arise from trialling the 3-option variant; we had some evidence from FCE that this variant of the task could perform consistently, although we did recognise that at this lower performance level the testing focuses being targeted could be subtly different.

Reflecting on the MM task, we felt, in a similar vein, it was worth investigating whether reducing the number of distractors from three to one, that is, a 6-option MM task rather than an 8-option one, could test to the level while making the task more accessible for students. This thinking reflected our concern mentioned above, to ensure the appropriate balance for the level in terms of the processing challenges posed by the task and the listening input.

Despite our initially positive feelings about our proposed changes, we reserved drawing any conclusions until substantial performance data from the field could be obtained. The observations we made from trialling these potential updates are described in a subsequent section in this article.

Introducing a new task to broaden the testing focus

We wished, if possible, to broaden the testing focus of the CAE Listening paper, so we considered adding a new task type. We felt the short extracts task offered the possibility of increasing the range of task types and test focuses of the paper. This task type has the potential to target the interaction between speakers (for example, the agreement of speakers); extending the range of text types in the sense that the task enables the inclusion of short conversational extracts from a variety of everyday contexts, clearly offering a different kind of interaction from the longer interview-type interactions found in the existing long-text multiple-choice Part 3.

We felt that introducing the short-extract multiple-choice task would also offer a benefit in terms of structuring the Main Suite exams more comparably, as this task is also used in different formats for the FCE and CPE Listening tests. We had two aims in trialling a version of this task at CAE: one, to see whether the task would be positively perceived by representative samples of the pre-exam candidature; and two, to investigate whether it could perform at the level according to our internal testing constraints, while also clearly discriminating candidate performance.

Writing and trialling the new and updated tasks

In terms of producing the proposed updated tasks and the new short extracts task, as far as possible the procedure mirrored that for the production of materials for the existing test specification. The Chair drew up initial guidelines for writing the tasks, experienced Item Writers were commissioned to write some tasks in the proposed formats which were then thoroughly pre-edited, edited, vetted and proofed. During the writing and editing process, we collected the observations of the Item Writers on writing the tasks and made notes on issues which arose during meetings, so that we could feed these into future considerations on the desirability and viability of the tasks. The updated tasks were then assembled into pretests, which were allocated to potential CAE candidates using the same mechanisms and international exam preparation population as would be the case with material destined for a live exam. The scope of the trialling population, as mentioned in the introduction to this article, resulted in a total sample of more than 700 prospective CAE candidates, which we felt was broadly representative of our international test-taking candidature for CAE.

Reviewing the trialled material

When it came to reviewing how the material had performed in the trials, we collated and analysed the performance data in order to produce descriptive and inferential statistics on the material, as well as gathering and reflecting on qualitative feedback in terms of both open and closed responses by candidates and teachers. As was the case throughout the review process, the Chair's input and observations at this review stage were key in directing and drawing conclusions on the proposed updates. In drawing conclusions on the quantitative and qualitative data on the proposed updates, we observed the following:

- There were statistical indications that the short extracts tasks we produced were performing at the target difficulty level of the exam, and discriminating clearly between more and less able candidates.
- The short extracts task was generating positive comments from candidates and teachers, who also indicated they thought it appropriate for the test.
- The 3-option multiple-choice variant of the long-text task (a proposed modification of the existing multiple-choice task) did not appear statistically to be meeting the target difficulty level of the exam; and the discrimination index tended to indicate a weak performance of this 3-option variant in terms of discriminating between more and less able candidates.
- Indications on the performance of the 6-option variant of the MM task appeared to be equivocal, suggesting that further research into the desirability of this modification would be merited.

So, while we felt the evidence was persuasive for the inclusion of the short extracts task as a new task which appeared to perform well at the level, and which had the potential to be popularly received by the test-taking and teaching populations, we were less convinced by the case for the proposed modification to the long-text task with 3-option multiple-choice questions. After further review, analysis and consideration of the data we obtained from trialling, we observed that it might be difficult to maintain the level of this 3-option variant over time. Of course, we did not want to introduce any changes which might diminish the current strengths of a tried-and-tested task, so we concluded there would be no obvious benefit in altering this task from its existing format. However, an interesting question arose at this point: if 3-option multiple-choice questions could be observed to function consistently on the short extracts task at this level, why did they not seem to do so in the context of the long-text task? Following further reflection and discussion, we made the observation that this may be partly due to the essentially different nature of the text types, in that the short extracts task may perform at the level because of the generally more intensely interactive nature of their source texts, and partly because of the wider range of testing focuses, including interactional focuses such as speaker agreement, which they target.

Further research into modifying the Multiple-Matching task

In order to investigate how candidates actually perceived the MM task, a research group was set up, consisting of Assessment staff and, in the role of research consultant, a senior member of the Research and Validation team. We conducted a small-scale qualitative research investigation (Khalifa 2006) using immediate retrospective protocol analysis with two versions of the task, on sample groups of prospective CAE candidates. There follows a brief description of this research, with a summary of the conclusions we drew.

The investigation consisted of two stages. Firstly, a focus group protocol was conducted following the trialling of a 6-option variant of the task and the existing 8-option variant on a sample of ten candidates which aimed to identify the key issues. Secondly, in order to gather further data on the issues identified by the first stage, a written questionnaire, based on the observations and analyses resulting from the focus group, was administered following the trialling of the two variants of the task on a sample of more than 50 candidates.

Following the focus group, it appeared the issue was not that the students had concerns about the current 8-option task format, in fact they seemed to think it was challenging and even stimulating, as CAE ought to be, but they did comment on how the tasks were written. They observed that it was not the number of options in the task which worried them, but that options longer than a phrase sometimes posed a challenge to read and keep in mind at the same time as listening. As the Part 4 consists of parallel tasks, to be completed simultaneously, they felt sometimes they were under too much pressure processing the options at the same time as they were listening to the extracts. In his report, the external consultant who conducted the focus group recommended, based on the observations recorded during the focus group, that the current established 8-option task be retained. Overall, we thought the candidate observations at this stage were extremely insightful, and in fact reflected issues raised in current psycholinguistic theory such as capacity theory, or the theory of working memory, which describes how there are constraints imposed on the amount of information which can be processed at any one time (Field 2004:326). We considered these student observations offered support for our growing consensus that there was not a clear need to modify the task, but that there was a need to continue to ensure that for the candidates, there should be a balance in terms of the work needed to process the task versus the challenge of listening.

When the second stage of the protocol analysis was conducted, the following data was gained:

- More candidates said they thought there was enough time to prepare for the 6-option rather than the 8-option variant.
- More candidates said they would prefer to do the 6-option task in the actual exam.
- The length of the options was felt to be an important factor.

Of course, we recognised that many extrinsic variables were present in this trial, such as the reaction of the members of the sample group to the content of the tasks which were selected for trialling. At this stage, from this limited research, it seemed that a majority of the candidates from the sample groups would prefer to do a CAE Listening test which had the proposed updated task type. However, after pretesting versions of the 6-option task on six separate groups each of around 100 candidates, in order to gain descriptive and inferential statistical data on a sample which were fairly representative of the test taker population, we could observe that statistically the 6-option task seemed to prove insufficiently challenging for the level, in addition to discriminating performance less clearly. Additionally, and importantly, we had a concern that changing the format from 8 to 6 options might, from the experience of writing the tasks and observations at both editing and review meetings, have the potential to focus the task more firmly on detailed listening, which would not accord with the construct in which we aimed to target a wide range of testing focuses, including listening for gist. So, having thoroughly explored this potential modification, we felt we would not be justified in changing the task from its existing 8-option format. However, a key observation from the research, and one which we felt should continue to be borne in mind, was that there should always be an appropriate balance in terms of the challenge of processing the task versus the challenge of the listening input.

Finalising the specification

At the end of the review period, following reflection on the discussions and analyses of the feedback from teachers and candidates and the performance data from trialling, an agreed fixed format for the CAE Listening test was finalised, for implementation from December 2008. Table 3 shows the finalised specification.

Overall, we feel that this final specification for CAE Listening comprises a broad range of tasks, text types and testing focuses, which offers the means to target most effectively a wide range of listening skills at the advanced level. In addition, we managed to achieve a time benefit for stakeholders in that we could reduce the overall time of the test by 5 minutes, a saving generated not by reducing the listening time, but by the reduction in the productive content and so transfer time required (candidates copy their responses to answer sheets at the end of the test). In terms of the sequencing of parts, given that we had set out in the review project to address all sources of feedback, and also to take the opportunity to ensure everything that could be done was done to make the exams accessible, motivating, and a positive test-taking experience for the candidates, we felt beginning the test with the short extracts task would be a desirable change. We felt candidates might feel more comfortable listening initially to a series of short interactional extracts rather than a longer, extended monologue. As a result of this consideration, we put the sentence completion task in Part 2. The existing 4-option MCQ long-text task follows in Part 3, and the existing 8-option variant of the MM task in Part 4. This new format for the CAE Listening test, which will go live in December 2008, now sits comfortably between the FCE and CPE Listening tests.

Table 3: Updated CAE Listening test specification December 2008 onwards

Part	Task Type	No. of Questions	Task Focus
1	3-option multiple-choice short extracts Heard twice	6	Feeling, attitude, opinion, purpose, functions, agreement, course of action, gist, detail etc.
2	Sentence-completion task Heard twice	8	Specific information, stated opinion.
3	4-option multiple-choice long-text task Heard twice	6	Attitude and opinion
4	Multiple-matching (2 parallel tasks) Heard twice	10	Gist, attitude, main points, interpreting context.

Test length: 30 items.
Test time: approximately 40 min (35 min listening, 5 min transfer time)

References and further reading

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

Field, J (2004) *Psycholinguistics: The Key Concepts*, Abingdon, Oxon: Routledge.

Khalifa, H (2006) *CAE Paper 4 Part 4 Modifications*, Cambridge ESOL internal report.

Reviewing Part 1 of the FCE Listening test

DIANA FRIED-BOOTH ESOL CONSULTANT

As part of the review of the FCE Listening test, consideration was given to all four parts of the test in terms of the task type, focus and format. In preliminary discussions, minor changes to Part 1 were explored; in particular, reading aloud of both rubrics and 3-option multiple-choice questions came under review since the delivery of these affects the timing not just of Part 1, but ultimately the overall length of the complete test. This brief report concerns the collection of feedback regarding this particular aspect of Part 1 of the paper.

In order to inform the decision-making process regarding, amongst other things, the presentation of rubrics in Part 1, candidates taking one of two FCE Listening pretests were asked to complete a questionnaire when they had finished the pretest. The questionnaire included YES/NO questions, as well as spaces for candidates to add their own comments about any aspect of all parts of the test they had just taken. The total number of respondents was 276.

The first question in relation to Part 1 required a YES/NO response:

> On the FCE Listening Paper, you *hear* the Part 1 questions as well as reading them. Do you like this?

Candidates were then able to add any comments about hearing the Part 1 questions. Out of the total number of respondents 79% answered YES, 13% NO and 8% left the question blank.

The second question also required a YES/NO response:

> Were the instructions clear in Part 1?

This question elicited a YES response of 94%, and 83% of respondents answered YES to the third question:

> Did you like Part 1?

The open-ended comments to each of these three questions were collated; few candidates added comments and there was no significant pattern to those comments which were given.

The candidate information, in conjunction with the feedback from teachers and consultation with other professionals, led to the decision that the Part 1 format would remain the same as from December 2008. There were, however, further considerations which informed this decision. It was recognised that if either the question rubrics and/or options were not read aloud, candidates would require additional reading time prior to hearing each extract in order to process the question.

It was also evident from the feedback that both candidates and teachers like the structure and content of Part 1. It has always been the intention that the short texts provide for a ‘fresh start’, allowing candidates to settle into the test knowing that if they miss an item, there are more opportunities to re-focus their attention. At the same time the short texts incorporate a range of mild accents, a variety of topics and vocabulary, and different task focuses involving both monologues and dialogues, all of which are perceived as providing a fair test of a candidate’s listening abilities.

The position of FCE in the Cambridge ESOL suite of examinations was another factor taken into account in support of retaining the Part 1 format. The student who progresses through the different levels will become familiar with the task types appropriate to each level. The PET Listening test provides visual as well as reading aloud support before the candidate listens to each short text; at CAE level only the opening contextualising statement is read aloud.

In conclusion, the FCE Part 1 sits comfortably between PET and CAE, forming part of the continuum that underpins the developing skills as learners move through the different levels.

Reviewing the FCE and CAE Speaking tests

CLARE HARRISON ASSESSMENT AND OPERATIONS GROUP

Introduction

Work began on the FCE and CAE Review Project in 2004. As with previous examination reviews, such as the CPE revision in 2002 (see Roger Hawkey’s article on page 2), this Project followed Cambridge ESOL’s standard procedures for reviewing examinations, which includes the following key stages:

- large-scale consultation with external stakeholders, such as candidates, teachers, school owners, examiners
- internal evaluation of existing tasks
- creation of draft specifications based on feedback and proposals from the previous two stages
- trialling of new tasks and analysis of findings
- communication of trial findings and resulting proposals to external stakeholders and elicitation of further feedback
- definition of final specifications.

This process is iterative and cyclical, allowing full consultation and analysis before final decisions are reached (cf. Figure 1 on page 4).

This article reports on some of the work undertaken to modify the FCE and CAE Speaking papers within the context of the FCE and CAE Review Project, concentrating specifically on Parts 2 and 3 of the tests.

Consultation phase 1

Two activities were carried out simultaneously to elicit opinion on the FCE and CAE Speaking tests, Chairs’ reports were commissioned and candidate and centre surveys were undertaken.

Chairs’ reports

One of these activities involved the Chairs of both FCE and CAE Speaking in producing a report summarising what they saw as the strengths of each paper and areas they felt could benefit from improvement. Overall, the Chairs concluded that the Speaking tests were effective in testing the constructs underpinning them since they elicit a variety of lexico-grammatical forms and provide opportunities for a variety of interaction types and patterns, e.g. individual, uninterrupted long turns and discussion. However, they considered that certain improvements could be made to ensure that the tests were accessible to all candidates.

An internal meeting was held to allow discussion of the contents of the Chairs' reports. Some of the key areas from this discussion are outlined below.

Part 2: individual long turn

Both papers currently contain similar Part 2 tasks. In FCE, each candidate is given a minute to compare and contrast two visuals while addressing a task, e.g.

> Compare and contrast these photographs, and say which of the people will remember this moment the longest.

CAE is broadly similar although the number of visuals may vary between two and five, and the task has two parts to it beyond the basic 'compare and contrast' aspect, e.g.

> Compare and contrast two or three of these pictures, saying how the atmosphere is different in each situation, and what the benefits of each method of learning might be.

For Part 2, the Chairs focused their attention on the difficulties that some FCE and CAE candidates have in addressing the task set. While all candidates fulfil the first part of the task in which they are required to 'compare and contrast', not all go on to satisfy the second part of the task. This was considered to be partly attributable to the fact that the task instructions are given orally, making them harder for some types of learner to process and remember. Another reason suggested was that instructions are given after the candidates receive the visual material, meaning that some candidates may not be paying full attention to the instructions because they are concentrating on the visuals.

One improvement to address this which was suggested by the Chairs was the inclusion of the task in written form on the picture sheet to enable visual learner types to see as well as hear it.

Part 3: collaborative task

Part 3 is also similar for both papers. Candidates are involved in a collaborative task based around picture prompts in which they are asked to discuss something together and to come to a negotiated conclusion, e.g.

> First, talk to each other about how interesting these different types of film would be. Then decide which two would be the best for students to discuss. (FCE)

> Talk to each other about what message these pictures communicate about the dictionary, and then decide which picture would be the best in appealing to people worldwide. (CAE)

The Chairs felt that the task type works well and provides appropriate opportunity for assessment of the intended testing focus (exchanging ideas, agreeing and disagreeing, interactive communication etc.), but as with Part 2, discussion centred on the difficulties some candidates have in addressing each part of the task.

Proposed improvements for Part 3 were: the inclusion of the task in written form on the picture sheet as for Part 2; and the creation of a 'split task' in which the Part 3 task is separated into two parts, allowing candidates two minutes to focus on the first part of the task and a further minute to focus on the second, e.g. for FCE:

> First, talk to each other about how interesting these different films would be. (2 minutes)
>
> Now, decide which two would be the best for students to discuss. (1 minute)

Candidate and centre surveys

At the same time as discussions were being held internally, opinions were also being canvassed externally by means of a large-scale centre and candidate survey. Feedback received in response to this focused on similar issues to those raised during the internal meeting. However, additional feedback was also received on the use of visuals as discussion prompts. Some respondents felt picture-based tasks may not elicit a sample of language appropriate to FCE/CAE level, and suggested that rather than having visuals as prompts, FCE and CAE Part 2 tasks could be based on a question and word prompts, such as are found in CPE, BEC, IELTS or Skills for Life Speaking[1].

Outcomes of the consultation

The two consultation activities indicated that it would be valuable to explore a number of avenues for Parts 2 and 3 of the Speaking tests. For example, it was suggested that, for Part 2, written task prompts could be added to the picture sheets. Trialling would then be carried out to see if candidates find it easier to address the task set when there is a written reminder of it for them to refer to. If the above were found to be effective, further trialling would be undertaken to determine the most appropriate form for the task prompt to take. Two forms were suggested: direct questions based on the task set or taken verbatim from the interlocutor's task. Trialling would also be undertaken to determine the most appropriate position on the picture sheet for the written-prompt to appear in.

Although the use of written-prompts, such as appear in CPE, would result in a change of testing focus for FCE and CAE Part 2 in that written-prompt tasks would not elicit language for comparing or describing, it was decided that Item Writers should be commissioned to produce such tasks in order to explore the possibility of their use with FCE and CAE candidates. Research should then be undertaken comparing the language elicited by the current picture tasks and the new written-prompt tasks.

It was decided that for Part 3, as for Part 2, written task prompts should be added to the picture sheet, and the effectiveness of this approach trialled as suggested above. Secondly, the Part 3 task should be trialled as a 'split task' with the two parts of the task separated to allow candidates to focus on one part at a time.

Task prompts on picture sheets

Picture sheets for Parts 2 and 3 were prepared to include a written replication of the task instructions. Trialling was undertaken with students preparing to take FCE and CAE, who carried out one current task, with instructions given solely orally, and one updated task, with written task prompts on the picture sheets alongside oral instructions. Candidates' performance was recorded for the purpose of

[1] For information on these and other English tests, please visit our website www.CambridgeESOL.org

analysis, and their opinions on the experience of doing the two tasks were elicited.

It was found that including written task prompts on the picture sheets increased candidates' confidence when undertaking the tasks as they felt that, if necessary, they could refer to the written-prompt if they were not able to remember the task that had been addressed to them. Although many candidates did not refer to the prompts, knowing they could do this was considered to be reassuring. Being able to refer back to the task wording was shown to help candidates focus fully on the task.

Follow-up activity 1

Once it had been established that written task prompts on the picture sheets helped candidates to address the task more fully, further trials were carried out to determine the most appropriate wording of these. Two forms of task prompt were trialled and compared. These were:

A) Task prompts consisting of an exact replication of the wording used by the interlocutor, which is delivered in the form of an indirect question. E.g. the CAE interlocutor says:

> I'd like you to compare and contrast two or three of these pictures, saying how the atmosphere is different in each situation and what the benefits of each method of learning might be.

and this is rendered as follows on the picture sheet:

- how the atmosphere is different in each situation
- what the benefits of each method of learning might be

B) Task prompts consisting of the interlocutor's words phrased as direct questions, e.g.:

- How is the atmosphere different in each situation?
- What might the benefits of each method of learning be?

Candidates referring back to the task prompts found it easier to do so when these were phrased as direct questions.

Follow-up activity 2

Task prompts were trialled in three positions on the picture sheet: at the top; in the middle; at the bottom. It was found that centrally placed prompts were distracting for candidates, and those located at the bottom of the page were easily overlooked. However, those placed at the top of the page proved to be positioned where candidates could see and refer to them without being obtrusive.

Part 2 written-prompt tasks: a new task type for FCE/CAE?

In response to feedback from some stakeholders which highlighted concerns that the Part 2 picture tasks do not elicit an appropriate level of language for FCE and CAE, it was decided that an alternative task type, based on that used for the long turn in CPE, should be trialled. Writers produced tasks for FCE and CAE which were based on a statement of opinion, a question and three short prompts, for example (at FCE level):

> It's difficult to keep fit and healthy nowadays. Do you agree?
> - eating healthy food
> - getting exercise
> - getting enough sleep

These were trialled with students preparing for FCE and CAE, who carried out both a current Part 2 picture task and a written-prompt task to allow comparison. The candidates were recorded and opinions elicited as before. In addition, two post-trial research studies were undertaken, summarised below.

Part 2 written-prompt vs. picture task: study 1

Members of the Inter-Varietal Applied Corpus Studies (IVACS) research centre at the University of Limerick, Ireland, were invited to undertake in-depth analysis of the data collected from the trialling in order to compare candidate output when doing written-prompt tasks in Part 2 with output when doing the current picture-based task.

Recordings of candidates doing written-prompt and picture-based tasks were transcribed by researchers. In each case, the same candidate carried out both a current picture-based task and an updated written-prompt task. Four transcripts of the written-prompt and four of the current picture-based tasks were analysed at each level. The same four students took part in each case, to allow comparability. Researchers were asked to subject transcriptions of candidate long turns to conversation analysis in order to answer the following research question:

> In the case of the Part 2 long turn, is there more evidence of coherence and cohesion in the candidate output when carrying out the written-prompt task than when carrying out the picture-based task?

The findings indicated that for both FCE and CAE Part 2 the written-prompt tasks encouraged candidates to produce more varied, coherent and complex language while the picture tasks seemed to produce less complex language and more hesitation. However, it should be noted that the picture-based tasks did not include the written question prompts on the picture sheets.

Part 2 written-prompt vs. picture task: study 2

Independent of the study described above, a marking project, based on the trial recordings, was also undertaken. This study was managed by Cambridge ESOL's Research and Validation Group and aimed to investigate the scores of eight candidates responding to the FCE and eight to the CAE Part 2 tasks with picture prompts and with written-prompts.

Each candidate was rated by 12 experienced Oral Examiners and scores recorded for Grammar and Vocabulary, Discourse Management and Pronunciation criteria (Interactive Communication was not considered relevant to Part 2 as it involves no interaction between the candidates). In addition, four of the examiners completed functional checklists to record the language functions elicited by the two task types. The research questions for this study were:

- Is there any difference in aspects of performance on each test as reflected in the marks awarded?
- Is there any difference in aspects of performance on each test as reflected in the functions elicited?

It was found that the CAE candidates in the study achieved a higher score on the picture-prompt tasks than they did on the written-prompt tasks while the FCE candidates' scores were fairly similar for both task types. The CAE candidates' better scores may have been due to their familiarity with the picture task, having very likely also prepared for it at FCE level.

In terms of functions, comparisons and descriptions were significantly more common with the picture-prompt tasks at both FCE and CAE levels and resulted in more speculation at CAE level. The CAE written-prompt tasks, on the other hand, elicited significantly more elaboration of opinions and reasons for assertions.

Part 3 'split' task vs. current task

At the start of the current Part 3 collaborative task, candidates are set two different conversational goals. They are expected to structure the interaction in order to address both of these. One of the proposals made by the Chairs in an attempt to help candidates address fully both parts of the Part 3 task was to split the task and its instructions into two parts. In the 'split' task, the two conversational goals are separated and sequenced by the interlocutor so that candidates receive only one task at the beginning of Part 3 which is then followed by the second task after they have discussed the first.

The IVACS research team was again invited to undertake research to compare candidate output when the instructions for Part 3 were given in two stages (the so-called 'split' task described above) with output on Part 3 in the current format. The research question which formed the basis of this study was:

> In the Part 3 collaborative task, are there differences in the language produced when candidates are given instructions in two stages as compared to when using the current instruction format?

The analysis of transcripts of candidates doing both tasks showed that although there was little difference in terms of the linguistic complexity of candidate output, there were differences in goal orientation and task completion. Researchers found that candidates in the sample either did not address both parts of the current Part 3 task, or did so only partially. It was found that candidates worked together, but they often did not achieve task completion. In the 'split' task, on the other hand, candidates tended to achieve both collaboration and task completion and it was found that separating the rubric into two parts provided candidates with two manageable tasks to concentrate on one at a time, allowing their discussion to be more task-focused. It should be noted, however, that task achievement per se is not directly assessed in the FCE and CAE Speaking tests, so these findings would not affect the assessment of this task, and that, as in the Part 2 study, the tasks were of the current type which do not include written question prompts on the picture sheets.

Consultation phase 2

Throughout the trialling phase, consultation with external stakeholders continued. As trialling findings were known and further proposals were made as a result, this information was presented at consultation meetings with key external stakeholders so that their opinion could be sought. With regard to the Speaking tests, some Chairs agreed that the issue of some candidates only partially addressing the Part 2 and Part 3 tasks was the one that they would most like to see addressed.

The option of basing the long turn in Part 2 on written-prompts elicited mixed opinions from stakeholders. Some felt strongly that the current picture tasks work very well and should be retained, maintaining that candidates like them and that they provide support for younger, or less sophisticated candidates, who may find it daunting to give a topical presentation such as is elicited by the written-prompt task. In addition, some of those consulted pointed out that apart from YLE Speaking tests, the picture-based tasks are unique to Main Suite within Cambridge ESOL and that for this reason it would be worth retaining them. However, those in favour of the written-prompt tasks felt that these were more authentic than the picture tasks and would allow candidates to produce more complex and coherent language.

The proposal to separate the task instructions for Part 3 also elicited a great deal of discussion. However, there was a minority of supporters for the 'split' task, who believed that in testing as in teaching, it is unwise to set two tasks at the same time since it is hard to focus on both at once. The majority of stakeholders were not keen on the 'split' task as they felt that it would make Part 3 disjointed, with the examiner intervening in the middle of the interaction to supply the second task. In addition, some of those consulted felt that knowing the goal of the whole task, which is only apparent on hearing the second part of the instruction, is essential to give a reason for doing the first part of the task. They felt that candidates could not carry out the discussion without knowing from the outset what the final goal is.

Most stakeholders agreed that providing task prompts on the picture sheets in both Part 2 and Part 3 would be a way to support candidates in fully addressing the tasks.

Final decisions

At the end of the consultation and research process, all the feedback and findings were taken into consideration to make final decisions about the content and format of the Speaking tests from December 2008.

Despite some of the findings from the small-scale research conducted by the University of Limerick into the written-prompt Part 2 task, it was finally decided that pictures should be retained at both FCE and CAE levels. The rationale for this decision takes into account the functions and topics of this task. The functions which form the basis of the testing focus of the long turn, i.e. comparing, describing, and in the case of CAE, speculating, are effectively elicited by the current picture task. These functions are not so effectively elicited by the written-prompts. Thus, replacing the picture tasks with written-prompts in Part 2 would have involved a considerable change to the underlying construct of one part of the Speaking test.

The topics that could be used to form the basis of the written-prompts are similar to those that are already in use in the Part 3 collaborative task, e.g. *lifestyle*, *the world around us*, *our working lives*. Had the written-prompts been adopted for Part 2, candidates could have found themselves talking about similar topics in two parts of the test and also making use of similar lexis and language functions to do this. Such linguistic and topic overlap would have reduced opportunities to elicit a varied sample of language for assessment. It would therefore have been necessary to re-think the focus and design of the collaborative task to avoid this potential overlap.

In the final analysis, introducing written-prompts for Part 2 appeared to necessitate a much greater change to the Speaking tests than the review process had shown to be necessary or desirable. Similarly, for Part 3, whilst there was evidence from the research conducted by the University of Limerick that indicated that the proposed 'split' task had some benefits over the current format in terms of goal orientation and encouraging candidates to address both parts of the task, there were other more compelling reasons to retain the current Part 3 format in the Speaking test. The proposed change to a 'split' task would have resulted in a greater focus on task achievement than is desirable since task achievement is not assessed in the Speaking tests and, furthermore, some stakeholder and internal feedback indicated strong resistance to the idea of separating the task into two parts.

In the face of these conflicting messages regarding Part 3, decision makers returned to the original issue raised by the Chairs of the papers and also by centre and candidate survey respondents, which focused on enabling candidates to fully address the tasks set, and determined on a course of action that would both address the issue and take account of stakeholders' views.

It was therefore decided that as the picture tasks were essentially popular and working well, they should be retained in both Parts 2 and 3. However, to address the issue of task focus in both parts, picture sheets should include task prompts, which candidates could refer to, allowing them to deal fully with the task set. Trialling had shown this approach to be successful and feedback from stakeholders regarding this proposal had been positive.

Conclusion

Although this review has resulted in only slight changes to the Speaking tests, the research carried out and opinions gathered on the subject have provided Cambridge ESOL with a wealth of valuable information. Clearly, monitoring and reviewing our tests is an iterative process and the insights afforded by the current review will act as a springboard for review projects of the future.

Developing revised assessment scales for Main Suite and BEC Speaking tests

EVELINA D GALACZI RESEARCH AND VALIDATION GROUP
ANGELA FFRENCH ASSESSMENT AND OPERATIONS GROUP

Introduction

It has become more and more the norm in the assessment community to devote time to the empirical validation of assessment scales prior to the scales being used in live conditions. Cambridge ESOL has always been committed to such validation prior to use, a point made in 1996 (Milanovic, Saville, Pollitt, & Cook) who used Rasch analysis in the development of rating scales for CASE, and followed up by Taylor (2000) in her overview of approaches to the revision of rating scales. The revision of the Main Suite/ BEC assessment scales for Speaking follows this tradition of *a priori* validation in performance assessment.

The assessment scales for Speaking for Main Suite and BEC were last updated prior to the introduction of the revised CPE examination in December 2002. At the time, it was agreed that a further review would take place once the new criteria had been used in live conditions, and the review of the FCE and CAE examinations provided an opportunity to do that.

Current assessment scales for Speaking

The 2002 assessment scales for Speaking follow a generic model, where the descriptors at each level are written in a similar way, but are interpreted at the level of the examination (see Weir & Milanovic 2003 chapter 7). At each level, 10 marks are available (0; 1.0; 1.5; 2.0; 2.5; 3.0; 3.5; 4.0; 4.5; 5.0), and descriptors are written for Bands 1.0, 3.0 and 5.0. While examiners have found these easy to use, analysis of data from live examinations has shown that little use has been made of the 5 marks available from 0 to 2.5, so truncating the 10 point scale. The descriptors of the 2002 scales include negative statements in all bands at all levels, and it was felt that this contributed to the under-use of marks 0 to 2.5. As a result, it was agreed to revise the criteria.

Using the Common European Framework of Reference (Council of Europe 2001; henceforth CEFR) as a starting point, the following guidelines were adopted to develop the descriptors:

- **Positiveness** – It was felt that positive formulation of descriptors should be attempted, if levels of proficiency were to serve as objectives rather than just an instrument for screening candidates.
- **Definiteness and Clarity** – Vague descriptors should be avoided since they can mask the fact that raters are interpreting them differently, and this makes the ratings less reliable. As far as possible, the descriptors should refer to concrete degrees of skill, i.e. what a candidate can be expected to do at a given level. To help make descriptors definite, concrete and transparent, a Glossary of terms should be introduced which defined each concept in specific terms with relevant examples.
- **Brevity** – Research has consistently shown that short descriptors are to be preferred to longer ones (CEFR, Appendix A) and that a descriptor which is longer than a two clause sentence can not realistically be referred to during the assessment process. An attempt should be made, therefore, to produce short, succinct descriptors.
- **Independence** – It was felt that each descriptor should have meaning without reference to any other descriptors, so each would be an independent criterion.

Revising the assessment scales for Speaking

In line with current thinking in the literature which advocates an empirically-based approach to rating scale construction (Council of Europe 2001, Fulcher 1996, Shohamy 1990, Upshur and Turner 1995), several scale development methodologies were followed in the design of the revised assessment scales: intuitive, qualitative and quantitative methods. Briefly, intuitive methods rely on expert judgement (ideally based on the literature) and the interpretation of experience. Qualitative methods involve interpretation of the information obtained, while quantitative methods rely on statistical analyses and the careful interpretation of results.

The full methodology supporting the current project took place in three phases, as described below.

Intuitive Phase (January–April 2006)

- reports from external experts, which included reviews of current ESOL practice in light of the literature and the experts' experience
- review of the reports by internal staff and an external reviewer and the setting out of design principles for the revised assessment scales
- production of Draft 1 descriptors.

Qualitative Phase (May–September 2006)

- a scaling exercise, which involved a rank ordering of the descriptors
- a verbal protocol trial, which involved raters' perception of the descriptors while rating performances
- an analysis of test performances at PET and FCE level by an external expert using a Conversation Analysis methodology. The aim was to identify discourse features associated with differently ranked performances and thus review the extent to which such features were captured by the scales
- production of Draft 2 descriptors.

Quantitative Phase (October 2006–April 2007)

- trial to assess the reliability of the revised criteria
- setting of "Gold Standard" marks. This involved using the revised criteria to assess performances on existing standardisation videos so that the marks could be used to standardise raters who would be involved in the marking of new standardisation videos
- further extended trial to confirm the soundness of the descriptors prior to the live roll-out.

In addition to the design principles outlined above, the descriptors for each level were mapped on to a common scale, so that, for example, the descriptors at A2 Band 5 were identical to those at B1 Band 3 and B2 Band 1. This was felt to be important since it suggested some rough equivalencies between different bands for different levels. There were, however, deviations from the 'stacking up' of levels: the descriptors for Pronunciation at levels C1 and C2 were identical, in line with current thinking on the assessment of Pronunciation (CEFR, Phonological Control Scale) and the descriptors for Grammar and Vocabulary were worded somewhat differently in the transition from B2 to C1, since at C1 they were divided into two separate assessment criteria ('Grammar/Vocabulary' at A2–B2 and 'Grammatical Resource' and 'Lexical Resource' at C1–C2).

In the remainder of the article, we will overview in more detail some of the research studies which were undertaken to inform the drafting of the analytical descriptors, namely the scaling of the descriptors, the verbal protocol study and the final marking trial.

Scaling of the descriptors

In order to explore the validity of the revised descriptors, draft descriptors were distributed to 31 Oral Examiners, who were divided into four groups. The participants were selected so that they represented all the levels in the Speaking Examiner framework: Senior Team Leaders, Regional Team Leaders, Team Leaders and Oral Examiners. It was felt important that the participating examiners should bring with them different levels of expertise and experience and so provide a more representative view on the revised descriptors. Each descriptor is sub-divided into different aspects of a criterion: altogether 64 sub-descriptors were identified.

Each group received a set of 20 of the 64 new descriptors and each participant was asked to match the descriptors to a test level on the Cambridge ESOL Common Scale. The relative 'difficulty' of the descriptors, based on examiner ratings, was estimated through FACETS. The examiner ratings were then compared to the levels intended by the scale developers. In addition, the consistency in the performance of examiners was investigated since lack of consistency might suggest difficulties in the interpretation of the descriptors.

Encouragingly for the validity of the revised scale, there was broad agreement between intended levels and examiner ratings. Descriptors placed at A1 by the developers were generally rated as the easiest by the examiners, and those placed at C2 were generally rated as the most difficult. There was some evidence that examiners were unwilling to use the extreme points of the scale with ratings clustering in the B1 to C1 range. In some cases the rank ordering did not match the anticipated level and in such cases the wording of the descriptors was looked into and revisited.

In terms of rater severity, the range fell between -2.41 and +2.58 logits. There was a significant (p‹.01, reliability of separation =.91) difference in harshness of examiners. This finding was in line with the available literature on performance assessment which indicates that rater variability is an inevitable part of the rating process and, as McNamara (1996:127) notes, 'a fact of life'. The results also indicated generally high levels of agreement between the raters involved. This was shown in the high point biserial correlations between the ratings made by each single examiner and by the rest of the examiners. The lowest point biserial correlation was at .63, which nonetheless suggested a high level of agreement between this individual's ratings and those of the other examiners.

In sum, this exercise showed that examiners were able to rank the draft descriptors in much the same order as intended by the scale developers. However, it was felt that there was a need to address through training the use of the full range of the scale since the consensus view of the descriptors resulted in a narrower clustering than was intended. There were also a number of issues of wording and clarification raised by the exercise which pointed to rewording or careful exemplification of some of the descriptors in the Glossary which accompanied the new scales.

Verbal Protocol Study

The guiding question in this study was, 'What do raters pay attention to when using the revised descriptors?' Similar to the study reported above, the participants ranged in terms of experience and expertise. Eight raters were asked to use the draft descriptors and award marks to a set of standardisation videos and in addition to fill in a questionnaire which asked for their comments on the usefulness of the revised descriptors and their experience using them.

In general, the comments focused on four themes:

The greater specificity, transparency, brevity and clarity of most of the descriptors

- 'More concise; it's easier to see the main points.'
- 'It is clearer what is expected at each band of each level, with less subjective interpretation.'

The need for greater clarity with some of the descriptors

- 'What constitutes "a good degree of control", "limited control" ?'
- 'The "range" aspect was quite difficult to judge.'

The greater ease of processing and applying the descriptors

- 'Much easier to process when marking.'
- 'Easier to apply perhaps just because they aren't quite so wordy.'

The use of positively worded descriptors

- 'I'm all in favour of positive statements at all levels – it is a matter of mindset.'
- 'The greater emphasis on positive achievement is welcome. One is encouraged to concentrate on what the candidate is capable of.'

The extended feedback which was received from the participants was a very rich source which informed the further revision of the descriptors for each criterion.

Marking trial

A large-scale trial was carried out after the descriptors were finalised in order to investigate the consistency and level of agreement of raters using them. The general aim was to confirm the soundness of the revised assessment scales prior to their being used in live conditions. In other words, it was important to provide evidence of the extent to which the assessment categories worked consistently at the five levels under investigation in terms of examiner severity, examiner agreement and misfit.

A total of 28 raters participated in this study, divided into seven groups. The raters provided a spread in terms of experience and position within the Cambridge ESOL Oral Examiner framework. In terms of location, most of the raters were based in Europe (in ten countries), one in Asia and one in Latin America.

A total of 48 test performances were rated, divided into five levels and eight exams: A2 (KET), B1 (PET and BEC Preliminary), B2 (FCE and BEC Vantage), C1 (CAE and BEC Higher), and C2 (CPE). The examinees had volunteered to participate in 'mock' speaking tests and had given consent to be video-recorded. Recordings took place at centres in the UK and Greece.

Each group of raters viewed a selection of test performances at different levels. The groups were constructed to ensure overlap between raters, levels and examinees. Multi-Facet Rasch Measurement (MFRM) was used as the method of analysis. The rationale for adopting this methodology rested on the widely accepted advantages of MFRM, such as the ability to provide estimates of the relative harshness and leniency of each examiner and to identify the least consistent examiners.

The findings indicated that there were different levels of rater severity. However, taking account of the view that rater variability is an inevitable part of the rating process, the important issue was how pronounced the differences in rater severity are. If the differences are within acceptable parameters, this would indicate that the raters are interpreting the scales in similar ways and, by extension, that the scales are performing at an acceptable level.

For practical purposes, Van Moere (2006) provides a range of -1.00 and +1.00 logits as being useful cut-off points of severity range. Applying these standards to the marks awarded in this trial, the majority of raters were found to be within acceptable parameters for harshness/leniency, indicating that they were following the expected standards and rating as an homogenous group. In addition, the majority of raters were internally consistent with the marks awarded. Only two examiners gave cause for concern as they were consistently too harsh and 'noisy' (i.e. showed too much unpredictability in their scores).

The acceptable range of examiner severity and levels of consistency for the majority of raters in this trial was seen as providing validity evidence for the revised assessment scales for Speaking.

Conclusion

The new set of assessment scales for Speaking for Main Suite and BEC examinations will be introduced in live conditions in December 2008. As shown in the above overview of the development and validation programme supporting the new scales, a great deal of time and effort has been devoted to their a priori validation. The triangulation of different methodologies (expert judgement; qualitative and quantitative studies) has engendered confidence in the scales and provided encouraging validity evidence for their use.

References and further reading

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

ffrench, A (2003) The change process at the paper level. Paper 5, Speaking, in Weir, C & Milanovic, M (Eds) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*, (Studies in Language Testing, volume 15), Cambridge: Cambridge University Press, 367–446.

Fulcher, G (1996) Does thick description lead to smart tests? A data-based approach to rating-scale construction, *Language Testing* 13(2), 208–238.

McNamara, T (1996) *Measuring Second Language Performance*, London: Longman.

Milanovic, M, Saville, N, Pollitt, A, & Cook, A (1996) Developing rating scales for CASE: Theoretical concerns and analyses, in Cumming, A and Berwick, R (Eds) *Validation in Language Testing*, Clevedon: Multilingual Matters, 15–38.

Shohamy, E (1990) Discourse analysis in language testing, *Annual Review of Applied Linguistics* 11, 115–128.

Taylor, L (2000) Approaches to rating scale revision, *Research Notes* 3, 14–16.

Upshur, J A & Turner, C (1995) Constructing rating scales for second language tests, *English Language Teaching Journal* 49(1), 3–12.

Van Moere, A (2006) Validity evidence in a university group oral test, *Language Testing* 23(4), 411–440.

Weir, C and Milanovic, M (Eds) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*, (Studies in Language Testing, volume 15), Cambridge: UCLES/ Cambridge University Press.

Overview of FCE and CAE Review Project research

FIONA BARKER RESEARCH AND VALIDATION GROUP
STEPHEN MCKENNA BUSINESS SUPPORT AND DEVELOPMENT GROUP
STEVE MURRAY ASSESSMENT AND OPERATIONS GROUP
IVANA VIDAKOVIC RESEARCH AND VALIDATION GROUP

Introduction

A wide range of activities have been undertaken from 2004 to the present day to inform the review of the FCE and CAE examinations. The updated specifications for both exams have benefited from over 40 research projects, reports and stakeholder surveys which have involved a range of internal and external personnel[1]. Most projects involved both inter-departmental collaboration within Cambridge ESOL and collaboration with external stakeholders and reflect the broad scale of this exam review. Whilst this issue of *Research Notes* has reported on some specific activities and outcomes in detail, there are many other projects which deserve mention, particularly those that are described in internal documents, therefore are unavailable for public dissemination.

It is important to recognise and acknowledge the extent and nature of inter-departmental collaboration that the research and consultation activities reported in this issue represent. All of the research projects represent the interaction of ideas and staff from different departments within Cambridge ESOL, including the Research & Validation, Assessment & Operations, Business Development & Business Management Groups – and the Customer Services Division – who worked together on the review of the FCE and CAE examinations. Externally, as has already been stated, thousands of people have been involved in a consultative or more specific capacity, for example the teachers and others who completed online stakeholder surveys and Chairs of individual papers who produced exam-specific reports. Everyone involved has made an important contribution and hopefully this issue reflects the diverse nature of the range of contributions made in the different areas of the FCE and CAE Review Project.

It should be noted that all of the research undertaken resulted in decisions that affected the review of FCE and CAE exams, either in the form of a specific action (i.e.

[1] Whilst too numerous to mention individually, we would like to express our sincere thanks to the internal staff and stakeholders who have contributed to the FCE and CAE Review Project. We hope that this issue celebrates the hard work and dedication of the many people involved in this extensive Project.

amending an existing task or mark scheme, or maintaining the current design of the exams) or a deepening of our understanding of the constructs underpinning these exams. All of the research and consultation undertaken therefore contributed to the production of new specifications and materials for the updated FCE and CAE exams, even where no specific changes were indicated by a particular report or the suggestions were subsequently not taken forward after trialling materials or further consultation and discussion.

In this article we offer an overview of the research and consultation activities which were undertaken during the review of the FCE and CAE examinations. We describe the areas which were explored and outline the topics of the numerous reports which were written by Cambridge ESOL staff and external consultants within the Review Project.

Stakeholder consultation

The Business Support and Development Group (BSDG) led the market research aspect of the FCE and CAE Review Project. During 2004, Cambridge ESOL conducted research to obtain feedback from stakeholders on possible changes to, primarily, the FCE and CAE examinations. This feedback informed the updates that were subsequently made to FCE and CAE, both in terms of the content of the tests themselves and of practical issues such as the overall duration of the exams.

The stakeholders were identified by Cambridge ESOL as being suitably qualified to offer an opinion and included candidates, Local Secretaries, teachers, examiners and supplier schools. Three versions of a stakeholder questionnaire were produced – for Candidates, Local Secretaries and an in-depth questionnaire for members of the ESOL teaching community. Questionnaires were completed online, although a paper version of the in-depth questionnaire was produced to augment the sample size. The respondents consisted of 1,900 candidates, 101 Local Secretaries from 20 countries and 625 detailed questionnaires were completed by teachers and examiners. The nationality breakdown of candidates completing the questionnaire broadly reflected the global ranking by candidature.

One particularly clear finding across all of the stakeholder groups consulted was that recognition of the exams (by universities and employers) is seen as the most important feature of an examination and that Cambridge ESOL examinations generally reflect this well.

In addition to the stakeholder surveys, Cambridge ESOL had face-to-face contact with twelve hundred stakeholders through a series of worldwide presentations, consultation meetings and invitational meetings. The consultation meetings, which were held in the UK, Spain, Germany, Italy, Poland, Switzerland, Romania, Argentina and Brazil were attended by over 450 teachers, directors of studies and other stakeholders. Delegate views were recorded through completion of detailed questionnaires and were then collated, and the results analysed. The findings were influential in leading to the final decision making process.

Overview reports and issue-specific projects

The overall review of each exam component was guided initially by an overview report by the Chair of each paper sponsored by the Assessment and Operations Group (AOG). These overview reports led to a range of issue-specific projects sponsored by AOG which are described in the following sections. The Research and Validation Group (R&V) led around twenty projects within the FCE and CAE Review and contributed to those led by other groups (most notably AOG). The aim of this Group's research program was to ascertain if there was a need for certain updates and to examine the impact of updates on the validity, reliability, impact and practicality (VRIP) of FCE and CAE exams. The research was undertaken in relation to three areas: the construct models of FCE and CAE, reviewing the mark schemes and assessment criteria, and investigating tasks, topics and general content within the exams.

The first group of projects aimed to define the underlying construct models of FCE and CAE, and then build and test these models. Since the construct is reflected in the structure of an exam, the main aim of these projects was to ascertain if there was a need to modify the structures of the exams. Specific projects looked into developing the construct model for each exam and the five different papers within each exam and the effect of fewer test items on reliability coefficients using a year's live performance statistics.

The second set of projects explored revising mark schemes for certain skills papers across FCE and CAE, sometimes for one exam, sometimes for both. Research included producing a new mark scheme for CAE Writing and producing revised assessment criteria for Speaking.

The final set of projects reviewed tasks within FCE and CAE and reported on trials using updated tasks or materials. The relevant research reports discuss the length of FCE and CAE Writing papers, reviewing the prompt type in FCE and CAE Speaking tests and multiple-matching task modification in the CAE Listening paper.

Other research projects undertaken by R&V and AOG staff are described in the following sections according to whether they investigated the overall exam construct or the separate language skills (Reading, Writing, Listening, Speaking) and language knowledge (Use of English). We finish by mentioning the time trialling studies which were done to ensure that, on a practical note, any modified tasks had been designed so that they could be done by candidates at the appropriate level in the time allocated.

Research into constructs

Cambridge ESOL devoted considerable time and resources to analysing and reflecting on the underlying construct of language proficiency in the FCE and CAE examinations, drawing widely on empirical evidence from exam data. Using various Exploratory Factor Analysis techniques Research and Validation staff built several plausible construct models for each paper and for the exam as a whole. The viability of each model was tested by Structural Equation Modelling (SEM) techniques (see Geranpayeh's article on page 8). This research additionally looked whether there might be support for the merging of any two papers.

Research was also undertaken into examining and defining the construct validity of FCE and CAE from a socio-cognitive perspective on overall language proficiency and the four language skills (see, for example, Weir and Shaw 2006).

Other research into contructs included:

- Investigating a proposal to report standardised scores on a common scale.
- Analysis of data gathered from live sessions on the nature of the candidature, including information such as candidate age, reason for taking the exam, etc.
- Widespread trialling on internationally representative groups of candidates to investigate the performance of the proposed task types for FCE and CAE. This extensive field research enabled the analysis of performance data in terms of both descriptive and inferential statistics on the tasks and items.

Research into specific skills

We now list many of the projects which researched one or more skills papers within FCE or CAE or both exams.

Reading

- Investigations were done into the suitability of particular task types for certain texts with particular emphasis on what types of text may be suitable for a productive task at FCE level.
- A report was written on the appropriacy of a productive reading task at the FCE level.
- Consideration of the proposal to introduce a sentence-completion task in FCE was carried out in a report which included a focus on how to write test items for the task and also outlined some issues in finalising the task keys.
- Another report on gapped texts discussed the advantages and disadvantages of both gapped sentence and gapped paragraph tasks. Several tasks of these types were analysed and commented on in detail.
- The advantages and disadvantages of adding short answer questions to FCE were discussed in an internal report.
- A report was commissioned into the appropriacy of using texts from fiction to test reading at CAE. Suitable sources of fictional texts were discussed as well as task-specific issues relating to texts, item types and test focus.

Writing

- A report was commissioned on the extent to which the element of persuasion may be distinctive in CAE Writing Part 1 tasks.
- Considerable resources were devoted to evaluating the impact of the proposal to shorten the FCE and CAE Writing test times. In the research, efforts particularly focused on the impact that reduced test time could have on the validity and reliability of the Writing Paper.
- Research was also undertaken to investigate the impact of reduced input and output length in Part 1 FCE and CAE tasks and scripts (see Cooze & Shaw's article on page 15).
- The CAE General Mark Scheme was analysed and a report was written containing proposals for a revised mark scheme. This was followed by a report which considered whether the revisions could improve the reliability and validity of the rating scale.

Use of English

- The appropriate lexical and structural balance of the paper was investigated, for example, investigations were made into whether the FCE and CAE papers could benefit from the inclusion of a productive test of vocabulary (the gapped-sentence task).
- Analysis of trialling data and further consultations generated observations on whether the CAE paper could benefit from the inclusion of a test of structure such as the key-word transformation task.

Listening

- Research was done into the issues surrounding finalising mark schemes for FCE productive tasks, to enable revised guidelines to be drawn up for the finalising of future mark schemes.
- Investigations were made and reports written concerning the appropriate range and balance of testing focuses for all parts in the tests.
- Verbal protocol analysis was done on CAE to investigate candidate perceptions about a particular task type and trialling was undertaken into the efficacy of the two variants of a task together with a focus group discussion (see Murray's article on page 19).

Speaking

- A report was written focusing on 'long turn' speaking tasks in FCE and CAE exams exploring the similarity and differences with the CPE model.
- The usability of FCE and CAE visuals as well as the accessibility and appropriacy of their topics were discussed in two further reports. A task-by-task analysis was provided, with sample tasks taken from FCE and CAE Speaking papers from 2003–4. The reports concluded with suggestions for certain improvements of this task type in the Speaking paper.
- Following a trial of changes to task types for FCE and CAE Speaking tests, a study was undertaken to investigate the scores of candidates responding to Part 2 in both FCE and CAE Speaking tests with a) a picture prompt and b) a written prompt (see Harrison's article on page 24).

Time trialling

The purpose of the time trialling observations was to provide data to inform the review process regarding the amount of time taken by candidates to complete the proposed updated tasks and components of the Reading, Writing, and Use of English papers. For each Reading and Use of English trial paper, grids, divided into five minute intervals, were completed to show which part of the paper candidates were working on at various points of the examination, and at what time they completed all of the tasks. For the Writing papers, we recorded the amount of time students spent preparing, and the time they started and finished writing each Part. Each trial paper was followed by a questionnaire which allowed students the opportunity to give feedback on the papers and tasks.

It was intended that the results of these observations, which were gathered from candidates in Brazil, Greece and multi-national centres in the UK, would give some insight into the way in which candidates approached the trial materials, how much time they required, and as a result, the potential implications of introducing the tasks to updated versions of the FCE and CAE examinations.

Conclusion

This overview has provided a snapshot of the breadth of research and consultation undertaken within the review of the FCE and CAE examinations. The dedication of thousands of people taking part in dozens of individual projects and consultation activities have made this a successful Review. The updated exams have benefits both for candidates in terms of clearer family resemblance of exam tasks and easier progression to more challenging exams within the Main Suite, and for the reliability and validity of the tests in terms of their construct, format and mark schemes.

Reference

Weir, C & Shaw, S (2006) Defining the constructs underpinning Main Suite Writing Tests: a socio-cognitive perspective, *Research Notes* 26, 9–14.

Conference reports

Cambridge ESOL staff attended and presented at various national and international conferences during the summer months, a selection of which are summarised below together with a review of a staff seminar given by Professor David Crystal in July.

LTRC 2007 – Barcelona, Spain

The 2007 Language Testing Research Colloquium (LTRC) took place in Barcelona in June, with considerable support and input from staff at Cambridge ESOL both prior to and during the event. This year's conference theme was *Exploring Diverse Methodologies and Conceptualizations in Language Testing Research*. Two of the pre-conference workshops had Cambridge ESOL involvement: Ardeshir Geranpayeh, in collaboration with Barbara Byrne, co-led a 2-day course in structural equation modeling, while Lynda Taylor ran a 1-day workshop on using qualitative research methods in language test development and validation. Nick Saville acted as coordinator and discussant for a symposium on current perspectives on language assessment for migration and citizenship, and Evelina Galaczi presented a paper on patterns of test taker interaction in the FCE Speaking test. Other ESOL staff attended the event for training and professional development purposes.

The conference was also an opportunity for several key awards to be made, including: the UCLES/ILTA Lifetime Achievement Award 2007, presented by Nick Saville (Cambridge ESOL's Director of Research & Validation Group) to Dr Charles Stansfield for his considerable contribution to the field and discipline of language testing and assessment; and the IELTS Masters Award 2006 (sponsored by the three IELTS partners), presented to Youn-Hee Kim of McGill University, Montreal for the Masters level thesis in English making the most significant contribution to the field of language testing. Details of Youn-Hee's dissertation were published in *Research Notes* 27 and her award was presented at LTRC by Dr Lynda Taylor (Cambridge ESOL) on behalf of the IELTS partners.

Nick Saville (Cambridge ESOL) presents the UCLES/ILTA Lifetime Achievement Award to Dr Charles Stansfield

IELTS Masters Award recipient Youn-Hee Kim with her supervisor Dr Carolyn E Turner (left) and IELTS partner representative Dr Lynda Taylor

EALTA 2007 – Sitges, Spain

The fourth annual EALTA conference was held in Sitges in June 2007. The conference consisted of a series of presentations on the theme *Good Practice in Language Testing and Assessment: Challenges and Praxis*. Jay Banerjee from the University of Lancaster gave a presentation entitled, *How do features of written language production interact at different performance levels?* Jay reported on a study which examined the defining characteristics of written texts at different points on a rating scale. Jay discussed data from IELTS Writing tests and exemplified the range of features such as lexical diversity, lexical sophistication, syntactic complexity, grammatical complexity and grammatical accuracy that were salient across bands on the rating scale.

Professor Cyril Weir's presentation in Sitges was based on work by Cambridge ESOL to articulate their approach to assessment in the skill area of Reading. The work builds on Cambridge ESOL's traditional approach to validating tests namely the VRIP approach where the concern was with Validity, Reliability, Impact and Practicality. It explores how the socio-cognitive validity framework described in Weir's *Language Testing and Validation: an evidence-based approach* (2005) might contribute to an enhanced validation framework for use with Cambridge examinations. In this presentation, Cyril focused on the cognitive processes involved in reading, particularly highlighting the processes that should be tested at C1 and C2 levels.

Vivienne May and Karen Ashton from Cambridge ESOL contributed a session entitled *"Small rewarding steps": Assessing learners and supporting teachers in UK classrooms*. Their focus was on Teacher Assessment qualifications for Asset Languages. They discussed the positive impact that the "small step" assessments have had on learners' motivation by providing both accessible and meaningful learning targets.

The last morning of the conference was dedicated to presentations and discussions on linking assessments to the Common European Framework of Reference (CEFR). The majority of presentations related to methodology and findings from studies using the Council of Europe's pilot *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages*. Jessica Wu and Rachel Wu talked about their experiences in Taiwan linking assessments to the CEFR while Richard Tannenbaum detailed standard setting methodology used to link TOEFL iBT to the CEFR.

BAAL/CUP Seminar, 18/19 June – University of Bristol

This seminar was the first to be organised following the creation of the BAAL Special Interest Group for Language Testing and Assessment and it benefited from the joint funding offered for such events by BAAL and Cambridge University Press. The theme of the seminar was *Language testing and assessment in applied linguistics: identifying reciprocity in language testing and applied linguistic research*. The event aimed to build on the growing presence of assessment and testing concerns with applied linguistics and to explore ways in which language assessment can contribute theoretical insights to second language acquisition research. The seminar provided a valuable opportunity to bring together researchers working at the interface of language testing and assessment research in different areas of applied linguistics and to critically review current thinking and practice from both sociocultural and psycholinguistic perspectives. A key aim was to contribute to a research agenda through dialogue between different communities in sub-fields within applied linguistics and language assessment.

Speakers included both established academics in the field and postgraduate students reporting on their current research. Among other papers, Lynda Taylor presented on the topic of *Are two heads better than one? Pair work in L2 learning and assessment*. This was an opportunity to profile some of the studies conducted by Cambridge ESOL over the past 15 years into the paired format speaking test and to consider how such research has enhanced our understanding of oral communication skills as a result.

KICE International Symposium – Seoul, Korea

Neil Jones presented at a symposium organised by KICE (Korea Institute of Curriculum and Evaluation) in Seoul in June 2007: *"Successful Language Education: Setting Standards" – An overview of the Common European Framework of Reference for Languages (CEFRL) and its relevance to Korean Language curriculum & evaluation*. The conference was supported by the British Council, Goethe Institut and the French Cultural Centre; the five invited speakers came from UK, Germany, France, Finland and Japan. Neil's presentation emphasised the close connection between Cambridge ESOL and the CEFR, and used the Asset Languages case study to stress that successfully implementing a proficiency framework concerns not only validity and reliability issues but important issues of impact.

Statistical workshops in Cambridge

Two 2-day workshops on *Statistical Analyses for Language Assessment* were organised by Cambridge ESOL's Research and Validation Group between June 18th and 21st. Professor Antony Kunnan, co-author of a book with the same title, was invited to conduct the workshop here in Cambridge. The purpose of the workshop was to introduce staff to basic and intermediate statistical analyses relevant for language assessment professionals. In line with the policy of broader cooperation with other parts of Cambridge Assessment (CA), the invitation was extended to colleagues in the ESOL Assessment and Operations (AOG) and CA Research Division (ARD) groups to participate. In addition to 14 validation officers, 5 research officers from ARD and 3 subject officers/managers from AOG attended the workshops.

Senior Team Leader conference – Cambridge

The recent Senior Team Leader (STL) conference was hosted by Cambridge ESOL in August 2007. Twenty two Senior Team Leaders attended this three-day event, which included sessions on themes of relevance to the STLs, such as new test developments, examiner standardisation and business updates. Sue Randall (Director, Business Planning & Marketing) opened the conference with an overview of

Cambridge ESOL's main business development strategies in its key regions and a look at current and projected performance. Other sessions explored a wide range of topics, including the Senior Team Leader role and its development into the future; market research; online examiner training and relevant Cambridge ESOL developments; the reconceptualisation of some Cambridge ESOL products and their suitability for young learners; the revised assessment scales for Main Suite and BEC; the updated FCE and CAE Speaking exams; current issues and research on paired tests.

By Hook or By Crook – David Crystal's linguistic journeys

David Crystal visited Cambridge ESOL in July 2007 to talk about the linguistic journeys described in his most recent book *By Hook or by Crook*. The title of the book is drawn from the discussion of the etymology of the phrase following a chance meeting with a shepherd in Wales. The fact that Crystal meets the shepherd while recording a radio programme for the BBC on accents, and that the shepherd in rural Wales turns out to have a strong Scottish accent set the pattern for the talk.

Crystal moved on in Wales to discuss the origins of place names and talked about the history of the longest place name in the United Kingdom, Llanfairpwllgwyngyllgogerychwyrndrobwllllantysiliogogogoch and the marketing background of the name. Crystal explained how in the nineteenth century, the town, which falls between Chester and Holyhead on a railway line needed a way to draw travellers, and trade, to the town. The ploy certainly worked as the town name now appears on postcards and souvenirs and pulls in tourists from all over the world. It would seem the use of place names for promotional uses is not limited to the United Kingdom as Crystal described being awarded a certificate for playing tennis in a small New Zealand town by the name of 'Wimbledon'. The sight of the road sign to Wimbledon had drawn him from his route to visit the place with the longest name place in New Zealand continuing to show the roving nature of the talk and the linguistic detours which he took in his travels.

Crystal went on to show how language does not always keep its originals associations as it travels. Lady Godiva is well-known from folklore in the Midlands of England as the woman who rode naked through the city of Coventry and psychologists now, understandably, link the name to exhibitionism. However, the name has apparently travelled to North American Universities and has links with various engineering faculties. It is unclear how this association has come about, as is the case with the connection to the exclusive chocolates which are now seen in shops in Europe.

David Crystal's talk, and the discussion which followed on the etymology and meanderings of language, certainly provided the audience with food for thought on both the historical aspects of language and on future routes which English is taking.

Crystal, D (2007) *By Hook or By Crook*, London: Harper Press.

Advance notice: Association of Language Testers in Europe (ALTE) 3rd International Conference, Cambridge 2008

The plenary speakers have been announced for the ALTE 3rd International Conference, which will take place in Cambridge from 10–12 April 2008. The theme of the conference is *The Social and Educational Impact of Language Assessment* with the following topics: *Language Assessment for Teaching and Learning, Language Assessment and Intercultural Dialogue* and *Language Assessment: Impact and Stakeholders*. This major multilingual event is open to all professionals with an interest in language assessment and associated issues.

Plenaries will be given by:

- Professor Micheline Chalhoub-Deville, *Standards-based assessment in the USA: Social and Educational Impact*
- Professor John A Hawkins, *Using Learner Language from Corpora to Profile Levels of Proficiency (CEFR) – Insights from the English Profile Project*
- Professor Tim McNamara, *Recognising the Other: Language assessment, immigration and citizenship*
- Dr Brian North, *The educational and social impact of the CEFR in Europe and beyond: a preliminary overview*
- Professor James E Purpura, *The Impact of Language Assessment on the Individual*
- Dr Lynda Taylor, *Setting language standards for teaching and assessment: a matter of principle, politics, or prejudice?*

Details about registering and booking accommodation for this event are available at www.alte.org/2008

There is also a competition to win a free place at ALTE 2008 by writing a paper on 'The Role of Assessment in a Multilingual World'. Further details are available at www.alte.org/2008/competitions

We look forward to welcoming you to Cambridge in 2008.