CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

Research Notes

Issue 62

November 2015

ISSN 1756-509X

CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

Research Notes

Issue 62 / November 2015
A quarterly publication reporting on learning, teaching and assessment

Printed in the United Kingdom by Canon Business Services

Research Notes

Issue 62

November 2015

Contents

Editorial

Welcome to issue 62 of *Research Notes*, a quarterly publication offering key insights and analysis of the Cambridge English approach to learning, teaching and assessment.

In 2015, two of the Cambridge English exams, *Cambridge English: First* and *Cambridge English: Advanced*, which for the sake of reference will be referred to as *First Certificate in English* (FCE) and the *Certificate in Advanced English* (CAE) throughout the issue, underwent a major review, in keeping with our commitment to continuously improve all of our products based on stakeholder feedback and current assessment theory. The conduct and outcome of this review is detailed in the introductory article by Ron Zeronis, the Assistant Director of the Cambridge English Assessment Group, and Ardeshir Geranpayeh, Head of Automated Assessment and Learning in Research and Thought Leadership.

Having set the scene with this broad overview, the following six articles reflect on a key component of the review, from discussions with the main test users to research underpinning the continuation of or changes to the assessment of each skill.

The first paper, by Howden and Mehta, focuses on the first stages of the review which began in 2011. Cambridge English Development Managers identified the Centre Exams Managers most familiar with FCE and CAE, and discussed the following key issues through an online questionnaire: why take the tests, what is their future, and should their format be revised? For both tests, improved prospects for employment and future study, coupled with their prestigious reputation, proved to be the crucial factors for choosing them. Stakeholders were certain that both tests should feature content suitable for study, work and general purposes, and recommended that CAE develop a more academic focus. Such findings allowed for further investigation of innovations such as combining the Reading and Use of English papers, and whether to shorten the length of the tests, both of which were eventually implemented.

The following paper, by Vidaković, Elliott and Sladden, delineates the rationale and execution of these revisions to the FCE and CAE Reading tests. The merging of the Reading and Use of English papers was recommended partly to give the Reading tests a family resemblance to other tests in the Cambridge English suite of exams, and also because it was justifiable from the construct perspective; reading comprehension models demonstrated the correlation between reading ability and language knowledge. The second revision, the shortening of the tests, was trialled in several centres across several countries. These trials showed that shortening maintained construct coverage and that most test items were within the acceptable range of difficulty, and led to refined test items such as the inter-textual reading task in CAE. These revisions improved the tests' focus and provided a basis for future test validation.

The Use of English tests are the focus of Docherty's paper. The author further discusses the correlation between reading ability and language knowledge, but also highlights the reasons for reporting Use of English separately. A breakdown of the pre- and post-revision formats of the papers shows how the components were brought in line with each other, and the results reported from the trialling of the shortened tasks show that the revisions were sensitive to the developing cognitive processes of candidates.

Although the changes to the other test components were not as significant as those to Reading and Use of English, some modifications were made to the Listening tests in light of the stakeholder input gathered over 2011–2013. Elliott and Chisholm describe how the range of topics in FCE Listening was changed to give stronger focus on the test's suitability for entry to Further Education. The concern that the CAE Listening test did not cover the C1 level construct adequately motivated an increase in items focusing on discourse representation, such as replacing formal interview recordings with items comprising two-way discussions.

Glasson and Galaczi's paper then takes us through the revisions to the CAE Speaking test, which focused on the context validity of the tasks and how they could better enable candidates to display their abilities. This included the inclusion of text-based prompts alongside visual prompts, and reducing the number of topics designed to elicit discussion, to avoid the superficial level of interaction that discussing several topics in a limited time might invite. The methodology and outcomes of the test trials are covered extensively; the uniform aim of these trials was to place the test taker at the centre.

Finally, Lim discusses the revisions to the tests of Writing. As with the Speaking test revisions, the tests were altered to allow candidates more room to display their abilities to the specifications set out by the Common European Framework of Reference (CEFR); for instance, both the word count and time allotted for the tasks were increased. Teacher feedback verified that the new versions were an improvement but adhered to the external framework of the CEFR. Feedback from students encouraged teachers to review their teaching practices, including allocating more discussion time to essay topics and developing students' planning and editing strategies.

The articles in this issue demonstrate the positive impact of the 2015 FCE and CAE test revisions. However, Cambridge English is aware that test revision is an ongoing process and further research will be undertaken to monitor statistical performance and construct coverage of these exams along with all of our other products.

Continuity and innovation: Updating FCE and CAE

RON ZERONIS ASSESSMENT AND OPERATIONS, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT
ARDESHIR GERANPAYEH RESEARCH AND THOUGHT LEADERSHIP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Introduction

In 2013 Cambridge English celebrated its first centenary of providing English language proficiency examinations for a wide range of purposes and educational contexts, both in Britain and over 142 countries worldwide. There have been continuous test evaluations and revisions in Cambridge English reflecting the changes in test theory, demographic changes in candidature, test users' attitudes towards test score use, the impact on various stakeholders and taking advantage of new innovations in technology. Many of the changes to Cambridge English examinations and the underlying constructs have been documented extensively in our *Studies in Language Testing* (SiLT) series (see for example volumes 15, 16, 17, 23, 26, 28, 29, 30, 35, 37, 38, 40 and 42). Central to all changes to our examinations is the concept of Continuity and Innovation. While the former ensures that the underlying constructs of the examinations remain relevant to the purpose of the test, the latter takes advantage of new methods and theories that allow us to provide more efficient and accessible tests to our stakeholders. In this special issue of *Research Notes* we report on the latest changes we have introduced in two of our flagship English language proficiency tests: FCE and CAE.

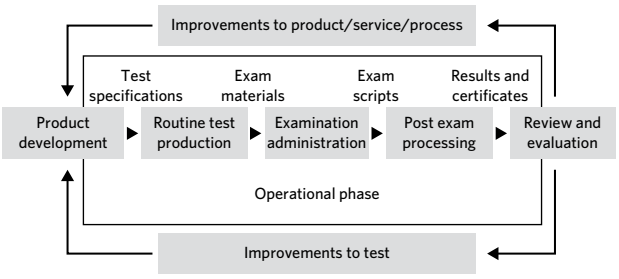
The *First Certificate in English* (FCE) is the second oldest qualification offered by Cambridge English Language Assessment. It was introduced by UCLES (as Cambridge English was then known) in 1939 as the *Lower Certificate in English*. The Lower Certificate was produced in response to a growing market need for an exam which tested proficiency at a lower and more functional level than UCLES' then flagship exam, the *Certificate of Proficiency in English* (CPE), as English started to take hold as the language of international commerce and travel. The exam underwent a major revision in 1975, when it was renamed the *First Certificate in English*. The 1975 revision was conducted against the backdrop of significant developments in the field of sociolinguistics, specifically the emergence of the concept of testing 'language in use', or language as used for practical communicative purposes. It was at this point that FCE began to more closely resemble the modern format used today (Falvey 2008:134–137). Following on from that landmark revision, further revisions to the format occurred in 1984, 1996, in 2008 and most recently in 2015. See Hawkey (2009) for the detailed history of the revisions.

Compared to FCE, the *Certificate in Advanced English* (CAE) is a relatively recent entrant to the Cambridge English exam portfolio. Introduced in 1991, it was designed to bridge the gap between FCE and CPE, and to fulfil a need by the market for a General English qualification at an advanced level which was suited for use for professional purposes. The aim of CAE was to 'offer a high-level . . . qualification . . . to those wishing to use English in their jobs' and 'to encourage the development of the skills required by students progressing towards CPE' (Hawkey 2009:113). CAE was a major innovation in test design and administration in 1991 and an essential component of the newly developed Cambridge English underlying psychometric common scale of language proficiency ladder which eventually paved the way for the adoption of the CEFR by Cambridge English as a means of reporting its scores. Like FCE, CAE has undergone a number of revisions to its format over the years, first being revised in 1999, then in 2008, and most recently in 2015 along with FCE.

The Cambridge English product revision cycle and the review of FCE and CAE

The regular review and evaluation of the exams in our portfolio is central to the quality assurance programme in place at Cambridge English. Product reviews allow us to ensure that all of our exams are up to date and remain fit for purpose. At the heart of our quality assurance programme is the Continual Improvement Cycle as shown in Figure 1 below:

Figure 1: The Cambridge English Continual Improvement Cycle (Cambridge English 2013:18)



We conduct annual performance reviews on all of our exams. These reviews consist of analysis of performance data from the previous year and are essentially a 'health check' to ensure the exams are performing as we expect them to. In addition, each exam undergoes an in-depth review every 5–7 years. The purpose of these reviews is to ensure the exams continue to meet the needs of the test users; that the exams continue to perform within their defined statistical parameters; and that they continue to reflect the most up to date testing methodology. The outcome of an in-depth review could be a recommendation to make no changes to the exam; to implement small changes as part of our routine production procedures; or to make major changes to the format and content of the exam. In some cases, external factors such as changes in the market or new strategic business objectives may prompt a more in-depth review of an exam.

When a decision is taken to make significant changes to an exam, a project to revise the test is initiated as part of the 'Product development' stage of the Continual Improvement

Cycle. Broadly speaking, a revision project will consist of three main stages: research and planning; design and development; and 'operationalisation' or the gearing up of internal systems and processes to run the new format. The review cycle is typically four years from initiation to launch of a revised exam. This is to allow for the extensive research needed to ensure the new test meets both market and assessment needs and to allow for the extensive trialling needed to fully test and validate the new design. It is also vitally important to give the market sufficient notification in advance of the new specification for any revised test in order to ensure that stakeholders, e.g. publishers of preparation material, teachers, etc. all have sufficient time to prepare for the changes.

The 2015 revision of FCE and CAE was conducted very much in line with our model of test development. In early 2011, it was decided to launch an in-depth review of both exams. Principal drivers for this included the launch of a new version of FCE, FCE for Schools, as well as the considerable time and effort Cambridge English was putting into increasing the recognition of CAE in the global Higher Education (HE) sector, a key business objective. The shift towards the use of CAE for the HE sector as well as the professional is influenced by the latest research in describing the C-levels in the CEFR. Green (2012:98) argues that many of the CEFR descriptors for C1 illustrate 'tendency towards academic and professional uses of language at the C levels'. This had to be reflected in updating the CAE. In addition, CPE (another C level exam) was itself in the process of a major revision which would lead to significant changes to the test format. It was against this backdrop, and in light of feedback on the exams being received from the market, that it was felt to be an opportune time to review FCE and CAE.

A project team was assembled consisting of members from key units across the business: the Assessment Unit; the Research and Validation Unit; the Business Development Group; the Operations Unit; and our Network Services Group. The review began in summer 2011 with an in-depth consultation and exam analysis phase. Key exam users were consulted, mainly via questionnaire, exam performance data was also scrutinised and the exam constructs analysed. Following this, a clear view formed that the exams should undergo significant revision. Once feedback from the initial research phase was considered, a draft revised specification was produced for each exam and approved by Cambridge English senior management for development. Material reflecting the new specification was commissioned and extensively trialled over a 15-month period, finishing at the end of 2012. As is standard in this type of revision, the development process was iterative – following each round of trialling, the results were analysed and adjustments to the specifications made. The final revised specifications were presented to stakeholders from autumn 2012. Although some trialling to fine tune some aspects of the tests was still being conducted into early 2013, the designs and content of revised FCE and CAE had been finalised and approved by the end of 2012. The operationalisation phase of the project started from early 2013. This included producing and pretesting a high volume of new test material; constructing new test versions for 2015; developing IT systems to administer and process the revised tests; training staff and examiners and informing our centre network; updating administrative procedures and documentation; intensive marketing and communication activity around the new specifications; and production of support material for teachers and students. This phase ended, and the project officially closed, with the first release of live results for each exam in early 2015.

Aims of the 2015 revision of FCE and CAE

The 2015 revision of FCE and CAE had a number of key aims:

For FCE, the aims were to ensure the revised exam:

- was suitable for use for Further Education (FE) study purposes
- was suitable for use for HE foundation or pathway courses
- was suitable for those who want to start working in an English-speaking environment
- retained coverage of all testing focuses
- reflected the most up to date methodological approach to communicative language testing
- was more user friendly in terms of its length.

For CAE, the aims were to ensure the revised exam:

- was suitable for use for HE study purposes
- was suitable for use for career advancement purposes
- retained coverage of all testing focuses
- reflected the most up to date methodological approach to communicative language testing
- was more user friendly in terms of its length.

In recent years Cambridge English has been working closely with educational institutions and government authorities to increase our presence in the education sector. We had identified a need for a better set of assessment tools than were typically being used by the market to measure at those CEFR levels most suitable for vocational and academic study, i.e. Levels B2–C2. In the 2011 review of FCE and CAE, when surveyed on the uses of the revised exams, stakeholders confirmed that the exams should be better suited to assess candidates' readiness for English-medium vocational and Higher Education courses (see Howden and Mehta, this issue).

Material for revised FCE was developed to ensure that the texts used in the test would have a more adult-level focus, and FCE now includes more topics set in the world of work and adult education which would be of interest to the target test takers. This has had the added benefit of helping to ensure there is a clear distinction for candidates in the content between FCE and FCE for Schools, and addresses an issue that had been raised by test users with the launch of FCE for Schools. In the revised specification, the latter continues to contain texts on topics of interest to, and within the realm of experience of, school-age test takers. CAE, meanwhile, was positioned more clearly as a General English exam with a strong academic flavour. The content is designed to better appeal to users who need the test for HE purposes, while retaining its appeal and usefulness to

the large number of candidates who continue to use CAE for career development and more general purposes. Many of the changes to the new CAE follow the recommendations of Khalifa and Weir (2009), Taylor (Ed) (2011) and Geranpayeh and Taylor (Eds) (2013) for improving the Reading, Speaking and Listening sections of the test. In addition to addressing test content in the context of test purpose, it was important to look at the test design for FCE and CAE in the context of Cambridge English's overall exam portfolio. In 2013, a revised version of CPE was launched which pointed the way to a new look for the higher level grouping of exams. Centres and candidates had long fed back that the exams were too long. This made them difficult and costly to administer and the perception was that candidate fatigue was sometimes an issue. The timing of the exam was addressed primarily by combining the Reading and Use of English components into a single paper. At the same time, a number of the tasks were shortened in length as the guiding principle was to make the exam as lean as possible without sacrificing either coverage or reliability. The test needed to retain the ability to provide in-depth measurement at C2 level and allow for the reporting of separate performance profiles of Reading and Use of English. The same approach was adopted for FCE and CAE. In both exams, the Reading and Use of English papers were combined into a single paper and tasks shortened where possible. Overlap of coverage was eliminated, which meant that some tasks that had featured in the 2008 version of the tests are no longer used in the 2015 versions (see Docherty, in this issue).

The construct review of both tests led to some further changes in the formats. In FCE and CAE, for example, it was decided to drop the 'set text' questions from the Writing papers. These questions had long been a feature of FCE and CAE and involved the setting of questions based on prescribed literary texts. Originally in place to encourage extensive reading and to promote good classroom practice, the inclusion of set texts in the Writing test was no longer seen as relevant to the target FCE and CAE test takers, supported by the fact that take-up of these questions was extremely low. However, it was decided to retain the set text option in FCE for Schools, where teachers clearly fed back that the texts supported their classroom practice and was beneficial for pupils in the schools context. In addition, although the CAE construct review reaffirmed the test's suitability for use for undergraduate-level academic purposes against the CEFR C1 academic descriptors, it did identify some areas where coverage could be improved. One of these was in coverage of more complex academic-style reading skills. To address this, following Khalifa and Weir's (2009) reading model, a new task type was developed to test the ability to understand different authors' opinions and stances on a topic across a number of different texts. Called 'cross-text multiple matching' this new task requires candidates to read four texts on the same subject by four different authors and to answer questions identifying the differences and similarities in the writers' views. This is a complex reading activity which mirrors the sort of reading an undergraduate might need to do when researching a topic or doing an assignment for their course. Similarly, new academic-style essays were introduced as compulsory tasks in the Writing papers of both FCE and CAE, replacing the transactional letters which had featured previously, while in the Speaking tests for both exams, the collaborative discussion task was revised to replace visual prompts with written prompts. This allows for the discussion of more abstract topics, particularly in CAE, and better allows candidates to display their ability to use more complex language. All of these enhancements have improved the effectiveness of the exams in measuring readiness for further and academic-level study.

Finally, in addressing the issues around test design discussed above, the task types used in the new formats were rationalised where possible so that the 'family resemblance' is now clearer than before across all three exams in the higher level grouping. The similarity in design, along with the alignment of the exams to the new Cambridge English Scale, and the reporting of results on this scale, mean that test users can see more clearly than ever before how the exams are linked and can much more easily see the distinction in level. This level of transparency means that candidates can clearly chart their progress up the levels as they become more proficient in using English.

The result of the 2015 revision is a set of exams in FCE, FCE for Schools and CAE which are better suited to meet the needs of the current test-taking population for those qualifications, better suited to meeting Cambridge English's strategic objectives and better able to meet the challenges of the coming decade.

References

Cambridge English (2013) *Principles of Good Practice*, Cambridge: Cambridge English Language Assessment.

Falvey, P (2008) English language examinations, in Raban, S (Ed) *Examining the World: A History of the Cambridge Local Examinations Syndicate*, Cambridge: Cambridge University Press, 131–157.

Geranpayeh, A and Taylor, L (Eds) (2013) *Examining Listening: Research and Practice in Assessing Second Language Listening*, Studies in Language Testing volume 35, Cambridge: UCLES/Cambridge University Press.

Green, A (2012) *Language Functions Revisited: Theoretical and Empirical Bases Across the Ability Range*, English Profile Studies volume 2, Cambridge: UCLES/Cambridge University Press.

Hawkey, R (2009) *Examining FCE and CAE: Key Issues and Recurring Themes in Developing the First Certificate in English and Certificate in Advanced English Exams*, Studies in Language Testing volume 28, Cambridge: UCLES/Cambridge University Press.

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.

Taylor, L (Ed) (2011) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press.

Stakeholder consultation: Review of FCE and CAE

DEBBIE HOWDEN BUSINESS AND MARKETING, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT
SANJANA MEHTA BUSINESS AND MARKETING, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Introduction

In 2011, an initial consultation with centres and teachers was undertaken to get their views on *First of Certificate in English* (FCE) and *Certificate in Advanced English* (CAE), particularly relating to what influences students' decisions to take either exam, what purposes the exams should be suitable for (study/career advancement/general travel and so forth), and also to get their opinions on combining the Use of English paper with the Reading paper to bring both exams in line with the revision of *Certificate of Proficiency in English* (CPE). This initial consultation formed an important part of the wider review to collect feedback from those who administer the exams and prepare students to take them, with a view to ensuring that the exams continue to be fit-for-purpose and meet the needs of users. The feedback received from these stakeholders, along with a review of exam performance data and of the exam constructs, informed a draft revised product specification which was then extensively trialled as part of our Cambridge English product revision cycle as mentioned in the introductory article (Zeronis and Geranpayeh, this issue). This research was similar to the consultation that informed the revision of CPE in 2009 (Docherty and Howden 2013).

Methodology

Cambridge English Development Managers were consulted to discuss centre perceptions and any feedback they had received about FCE and CAE since the previous revision in 2008. Development Managers are in regular contact with Cambridge English exam centres as part of their role to develop and grow new and existing business. Their feedback was used to inform the design of the surveys for teachers and centres.

In June 2011, individuals who had subscribed to the Cambridge English Teacher support website were emailed a link to an online survey for teachers. Simultaneously, approximately 800 centres which ran either FCE or CAE were invited to participate in a separate online survey for centres.

Responses were received from 2,053 teachers in 11 countries. Almost half of the teachers (47.3%) taught in a private language school, while just over a quarter of respondents (26.4%) taught in state-funded schools. Three quarters (74.8%) had prepared students for FCE and almost half (49.9%) had prepared students for CAE in the three years prior to this consultation (from 2008 onwards). It was this grouping within the total sample that was considered to be the most familiar with the tests and they were asked specifically to comment on:

- what influences students' decisions to take the tests
- the future purpose and content of the tests, and
- their views on which papers could be combined.

Responses were also received from 213 centres in 12 countries. Of the total respondent centres, 180 (84.5%) had administered FCE and 174 (81.7%) had administered CAE between 2008 and 2011.

Results

Factors influencing students' decision to take the exams

Table 1 shows that the vast majority of teachers and centres believed that the two most influential factors for students to choose FCE were: that 'it improves their job prospects' (91.4% teachers, 90.0% centres, likely or very likely) and that 'it is recognised for further study/training purposes' (88.5% teachers, 88.3% centres, likely or very likely). Teachers also indicated that their recommendation or the recommendation of their institution were important influencing factors (81.6% selected 'likely or very likely' for the response option 'it is recommended by their teacher/institution').

Table 2 shows that the vast majority of teachers (93.2%) and centres (92.5%) believed that 'it improves their job prospects' was a key factor influencing their students' decision to take CAE. Furthermore, teachers (85.7%) and centres (90.8%) felt that the recognition of CAE for admission to universities or colleges in English-speaking countries was also a primary influencer.

The prestige of both exams (selected by approximately 80.0% of teachers and centres) was considered to be an important factor influencing students' decision-making.

Of the response options given in the survey to understand factors which play a role in students' decision-making, the option 'the exam dates are convenient' was selected by less than 50.0% of the respondents, showing that this was not one of the most important determining factors in the selection of FCE or CAE.

The survey also provided the respondents with the opportunity to comment on any other factors influencing

Table 1: Main factors influencing students' decision to take FCE

Likely and very likely influencers	Teachers	Centres
It improves their job prospects	91.4%	90.0%
It is recognised for further study/training purposes	88.5%	88.3%
It is recommended by their teacher/institution	81.6%	78.3%
It is a prestigious exam	80.2%	82.8%
It is for their personal development	67.3%	72.8%
It is recognised as a school-leaving qualification	63.8%	62.2%
It is recommended by their parents	58.0%	58.9%
It can be taken at a nearby centre	56.7%	70.0%
It is recommended by their friends	52.5%	67.2%
The exam dates are convenient	37.5%	46.1%

Base: 1,535 teachers and 180 centres that had prepared students for, or administered, FCE between 2008 and 2011. This is the base for all responses to questions on FCE.

Table 2: Main factors influencing students' decision to take CAE

Likely and very likely influencers	Teachers	Centres
It improves their job prospects	93.2%	92.5%
It is recognised for admission to universities or colleges in English-speaking countries	85.7%	90.8%
It is a prestigious exam	79.4%	82.8%
It is recommended by their teacher/institution	79.4%	78.8%
It is for their personal development	75.2%	74.1%
It is recognised for admission to universities or colleges locally	64.9%	74.1%
It can be taken at a nearby centre	56.2%	68.4%
It is recommended by their parents	51.1%	51.8%
It is recommended by their friends	49.6%	58.6%
The exam dates are convenient	36.4%	47.7%

Base: 1,025 teachers and 174 centres that had prepared students for, or administered, CAE between 2008 and 2011. This is the base for all responses to questions on CAE.

students' decision to take the exams. There was mention of the exams being incorporated in the school curriculum by a few respondents, illustrated in the following quotations:

> As with FCE, the CAE has become part of the way that English is being taught in all levels of German schools and those who have CAE, in particular, are able to work and study abroad as well as work in the increasing number of companies who use English as their everyday language of communication here in Germany. (Centre – Germany)

> Course at our college are fully tailored to the FCE qualification. (Centre – UK)

Many other comments supported the view that the recognition and prestige of FCE and CAE was an important influence on students:

> Many candidates are required to take FCE by their colleges and universities in order to enrol for certain courses or modules. (Centre – Germany)

> Local prestige, worth 'credits' at local universities, focus of private language centres on Cambridge English exams, in preparation for CAE later on (Erasmus, exchange study programmes, required leaving level at several private further education centres). (Centre – Spain)

> The results can certify at an international level their efforts and dedication of years of study of the English language and assure them that they have taken a course at a top quality institution. Having a Cambridge English FCE certificate to prove their proficiency level in the language almost becomes a matter of status and gains respect in their curriculum. (Centre – Brazil)

Content for the future

When asked to select what content is 'important or very important' to be included, more than three quarters of centres and teachers wanted to ensure that both FCE and CAE included content suitable for all of the suggested purposes: content suitable for general purposes, study purposes, and also for career advancement – as illustrated in Tables 3 and 4. However, the order of importance differed between teachers and centres for FCE as teachers rated that having content suitable for general and study purposes was more important, whilst centres rated that having content suitable for general and career advancement was most important. Teachers and centres did however agree the order of importance for the content for CAE. A few centres also raised the issue that the topics for FCE should be more age appropriate, particularly following the launch of FCE for Schools:

Table 3: Important and very important content to include in the future – FCE

Type of content	Teachers	Centres
Content suitable for general purposes e.g. travel, personal interest	87.8%	81.1%
Content suitable for study purposes	81.5%	75.6%
Content suitable for career advancement	75.1%	77.2%

Table 4: Important and very important content to include in the future – CAE

Type of content	Teachers	Centres
Content suitable for study purposes	85.7%	85.1%
Content suitable for career advancement	84.6%	83.4%
Content suitable for general purposes e.g. travel, personal interest	83.0%	78.2%

> Now that there is FCE for Schools, FCE can become a truly 'adult' exam with content more appropriate to study and work contexts. (Centre – Italy)

> The fact that we can count on an ordinary FCE and one version 'for school' has turned the exam into a very convenient option given our students' age and maturity. (Centre – Argentina)

Some teachers also suggested an academic focus for CAE in the future:

> Academic topics for the reading and writing part as well for the oral exams. (Teacher FCE and CAE – Germany)

> The CAE should reflect much more that many candidates are students at higher education. What about having a general CAE AND an academic CAE, like IELTS? Otherwise, the CAE should become more academic in focus; writing a story is not particularly relevant for university students, but writing a report based on data, or an academic essay, summary or similar academic text would provide better preparation for their university studies. Further the reading test should cover less generally themed texts and concentrate more on texts that students would be more likely to come across, such as in textbooks, etc. (Teacher FCE and CAE – Italy)

> As in FCE – general and academic paper might be useful for employment purpose/further education. Please maintain the Use of English section, that is what makes it different from other exams and at this level it really shows the mastery of the language and understanding of the nuances. (Teacher FCE and CAE – UK)

> CAE may benefit from becoming a more widely recognized academic exam or possibly split into two different types of exam – one for academic purposes and one for more professional needs (similar to the IELTS split) as might the CPE. (Teacher FCE and CAE – Italy)

Reducing the format from five papers to four

As shown in Tables 5 and 6, the initial findings indicated that centres and teachers were broadly satisfied with the existing five-paper format of the test, but their second preference would be for the Reading and Use of English papers to be combined, which would bring the papers in line with the changes that had been made to the revised CPE. This topic was explored further as part of the second phase of the consultation with stakeholders, where it was discovered that combining the Use of English paper with the Reading paper was more favourably

received once it was explained that the reporting would continue to show results for the Use of English component.

Table 5: Which of the FCE papers could be combined?

Type of paper	Teachers	Centres
Reading and Writing	17.2%	13.3%
Reading and Use of English	20.1%	28.3%
Writing and Use of English	14.2%	16.7%
A different combination	1.4%	1.1%
No combination, retain the five-paper format	47.0%	40.6%

Table 6: Which of the CAE papers could be combined?

Type of paper	Teachers	Centres
Reading and Writing	14.6%	12.6%
Reading and Use of English	18.2%	24.7%
Writing and Use of English	11.7%	15.5%
A different combination	1.1%	1.1%
No combination, retain the five-paper format	54.3%	46.0%

Shortening the length of the exams

Shortening the length of the exams was supported by more than a third of centres and teachers, as shown in Table 7. In the second phase of the consultation this was explored further.

Table 7: How important is it for the updated exam to be shorter in length?

Exam	Teachers	Centres
FCE	36.5%	35.5%
CAE	38.8%	42.0%

Any concerns raised around shortening the length of the exam focused on ensuring that the changes should not have any impact on the quality of results and on the prestige of the exams, as illustrated in the following quotations:

> If it was to be decided to shorten the exam, make sure that the standard is the same. (Centre – Italy)

> Ensure that its credibility is maintained: there is a danger that by shortening the exam [FCE] it will be perceived as being less rigorous. (Teacher FCE and CAE – Spain)

> It is the length of the exam which makes it perceived as a serious exam. To shorten it would undermine its prestige. (Centre – Italy)

Other considerations

When asked about what else should be considered when these exams are being revised, some centres and teachers expressed their support for bringing the format of the exams in line with each other and in line with the changes being made to CPE:

> Make format of FCE and CAE identical to avoid confusion for students and teachers alike and make teaching easier especially in smaller institutions where FCE and CAE students may be taught in the same class . . . (Teacher FCE and CAE – UK)

> Consistency with other levels (FCE and CAE). (Teacher CAE – Poland)

> Must consider the length in relation to the other [higher level] examinations. As the CPE is being revised and it appears that it will be shorter than the CAE, this will need to be explained as it is not logical to test takers and teachers. (Centre – Switzerland)

Conclusions

This initial consultation showed the importance that recognition of the exams for further study and to improve job prospects plays in encouraging learners to take FCE and CAE and was also reflected in the need to ensure that the content of the exams is suitable for general and study purposes and career advancement. The findings informed the second phase of the consultation where it was possible to explore in more detail how the format of the exams should be revised in future.

Consulting with our stakeholders in this way continues to be an important stage of our product review process as it provides us with an opportunity to find out first-hand from key stakeholders what should be taken into consideration when revising the exams and guides test development specifications for further exploration. This research had a direct impact on the revisions made to FCE and CAE.

References

Docherty, C and Howden, D (2013) Consulting stakeholders as part of the Cambridge English: Proficiency exam revision, *Research Notes* 51, 18–21.

Revising FCE and CAE Reading tests

IVANA VIDAKOVIĆ RESEARCH AND THOUGHT LEADERSHIP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT
MARK ELLIOTT VALIDATION AND DATA SERVICES, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT
JULIE SLADDEN ASSESSMENT AND OPERATIONS, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Introduction

The *First Certificate in English* (FCE) and *Certificate in Advanced English* (CAE) Reading papers were revised to refine test constructs and ensure they remain fit for purpose. As part of these goals, it was necessary to ensure that FCE and CAE Reading tests were more appropriate for those wishing to study in an English-medium Further Education and university

environment, respectively. Reading and Use of English were combined into a single paper, which required shortening some tasks, while maintaining the current levels and coverage. These revisions were in line with changes introduced in the *Certificate of Proficiency in English* (CPE), maintaining identity and a coherence to the set of exams. Here, we focus on reading comprehension tasks, while Use of English is discussed in detail by Docherty (this issue).

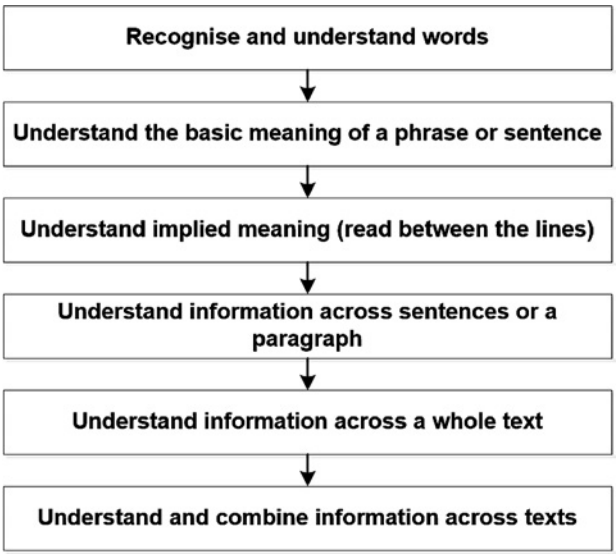
FCE and CAE Reading and Use of English: Two sides of a coin

The key driver for the revisions to FCE and CAE, that of bolstering their suitability for use as entrance requirements for Further and Higher Education respectively, requires examination of the construct coverage in terms of both the cognitive process and sub-skills required and the contextual features of the tests such as text types and linguistic complexity in relation to the demands of the context of use. This process, and how it resulted in specific changes to the CAE Reading paper, are discussed below.

On a surface level, however, the most obvious change to FCE and CAE Reading and Use of English papers consisted of merging the two into a single paper, as had been previously done in CPE (see Zeronis and Elliott 2013:23). This helped maintain a family resemblance in the examination suite, but the merge is also justifiable from the construct perspective. For example, lexical and grammatical knowledge enables and correlates with reading ability, which is discussed by Docherty (this issue). In this section, we focus on how the merge fits into the reading comprehension model which underpins Cambridge English language exams.

Even though the focus of Use of English tasks is on lexical and grammatical knowledge, they, along with Reading tasks, require reading comprehension. All are text based, apart from the sentence-based key word transformation task. Together, Reading and Use of English tasks activate a wide range of cognitive processes and reading types represented in the model of reading comprehension in Khalifa and Weir (2009). The model is based on evidence-based research into reading in one's mother tongue, and shows that cognitive (or mental) processes activated during reading can be lower level (at the level of word, phrase and sentence) and higher level (across sentences, at the level of paragraph or across paragraphs and at the level of the entire text or across texts) (see Figure 1).

Figure 1: A model of reading comprehension (adapted from Khalifa and Weir 2009:43)



It is generally accepted that both low- and high-level processes (can) happen simultaneously (Williams and Moran 1989, Khalifa and Weir 2009), but it is also possible that some tasks activate one or the other more predominantly. For example, Use of English tasks primarily test a candidate's ability to use words, phrases and grammar, and the type of reading required is typically at the level of word, phrase or a sentence due to the narrow, lexico-grammatical, task focus and/or due to task format (e.g. gapped sentences) (see Table 1 and Table 2). On the other hand, Reading tasks primarily require understanding of the main idea, detail, text organisation, implication, attitude, opinion, etc. With their broader task focus and longer passages, they require the ability to read and understand across sentences, paragraphs, a whole text or several texts (see Table 1 and Table 2). Therefore, Reading and Use of English tasks together cover a broad range of reading skills employed by a fluent reader in everyday life.

Table 1: The revised FCE Reading and Use of English paper

FCE Reading and Use of English	Test part (items; marks)	Task type	Task focus and reading skills required
Tasks contributing to the Use of English score	Parts 2–4 (22 items in total; 28 marks)	Open cloze, Word formation, Key word transformation	Task focus: Vocabulary and grammar Reading skills: Reading comprehension at word, phrase and sentence level
Tasks contributing to the Reading score	Part 1 (8 items; 8 marks)	Multiple-choice cloze	Task focus: Vocabulary and grammar Reading skills: Comprehension at word/phrasal/sentence level and across sentences
	Part 5 (6 items; 12 marks)	Multiple choice	Task focus: Detail, opinion, attitude, tone, purpose, main idea, gist, meaning from context, implication, text organisation features (exemplification, reference) Reading skills: Reading comprehension across sentences/a paragraph
	Part 6 (6 items; 12 marks)	A gapped text with the removed sentences placed in jumbled order after the text	Task focus: Cohesion, coherence, text structure Reading skills: Reading comprehension within and across paragraphs and also across a whole text
	Part 7 (10 items; 10 marks)	Multiple matching with one long text or up to six shorter texts	Task focus: Detail, opinion, specific information, implication Reading skills: Reading comprehension across sentences, across a paragraph or paragraphs

Table 2: The revised CAE Reading and Use of English paper

CAE Reading and Use of English	Test part (items; marks)	Task type	Task focus and reading skills required
Tasks contributing to the Use of English score	Parts 2–4 (22 items in total; 28 marks)	Open cloze, Word formation, Key word transformation	Task focus: Vocabulary and grammar Reading skills: Reading comprehension at word, phrase and sentence level
Tasks contributing to the Reading score	Part 1 (8 items; 8 marks)	Multiple-choice cloze	Task focus: Vocabulary and grammar Reading skills: Comprehension at word/phrasal/sentence level and across sentences
	Part 5 (6 items; 12 marks)	Multiple choice	Task focus: Detail, opinion, attitude, tone, purpose, main idea, implication, text organisation features (exemplification, comparison, reference) Reading skills: Reading comprehension across sentences/a paragraph
	Part 6 (4 items; 8 marks)	Cross-text multiple matching with four short texts (see the Appendix)	Task focus: Understanding of opinion and attitude; comparing and contrasting of opinions and attitudes Reading skills: Reading across texts
	Part 7 (6 items; 12 marks)	A gapped text with the removed paragraphs placed in jumbled order after the text	Task focus: Cohesion, coherence, text structure, global meaning Reading skills: Reading comprehension within and across paragraphs and also across a whole text
	Part 8 (10 items; 10 marks)	Multiple matching with one long text or several shorter texts	Task focus: Detail, opinion, attitude, specific information Reading skills: Reading comprehension across sentences, across a paragraph or paragraphs

The merging of Reading and Use of English papers has not changed the constructs of the previously separate papers, aside from the fact that the scores on one task which was previously part of the Use of English paper – the multiple-choice cloze with a vocabulary focus – now contributes to the Reading score rather than the Use of English score (see Docherty in this issue; see also Table 1 and Table 2). It is still possible to clearly distinguish Use of English tasks from Reading comprehension tasks, in terms of task focus, task format and the levels of cognitive processes they predominantly activate. In view of the different task focuses and aspects of language ability that Reading and Use of English tasks tap into, scores for Reading and Use of English are reported separately in statements of results and certificates. Reading, Use of English, Listening, Speaking and Writing scores are then aggregated to arrive at the total score for each candidate.

The merge of the two papers was made possible by reviewing existing tasks and identifying areas of overlap. By shortening some tasks and excluding others, the overlap was removed, while maintaining the coverage of the previous version. As a combined paper, the components now take less time to complete than was the case when they were separate papers. With Reading and Use of English now lasting 1 hour and 30 minutes in CAE and 1 hour and 15 minutes in FCE, administering the two as a single paper was deemed to be more efficient and also in line with the previous changes to CPE. The shortening of the pre-revision tasks and their trialling are discussed next.

The shortening of FCE and CAE Reading tasks

During the revision, it was decided to retain all pre-revision FCE, and most pre-revision CAE Reading tasks, but in shortened formats, in order to maintain construct coverage. Most texts were reduced by 50–200 words, and the tasks were reduced by 1–5 items (see Table 3). In addition, one CAE Reading task was excluded due to the construct overlap with another text (see *Refining Reading test constructs*).

Table 3: A comparison of pre- and post-revision FCE and CAE Reading tasks

Reading comprehension tasks	Words/ items	Pre-revision FCE	Post-revision FCE	Pre-revision CAE	Post-revision CAE
Multiple-choice task	Range of words	600–700	550–650	600–850	650–750
	Number of items	8	6	7	6
Gapped-text task	Range of words	550–650	500–600	650–800	800–900
	Number of items	7	6	6	6
Multiple-matching task	Range of words	600–700	500–600	650–800	600–700
	Number of items	15	10	15	10

A trial was carried out to determine if the shorter tasks perform well statistically and if they remain appropriate for the levels of the two examinations. Reading and Use of English tasks were administered and trialled together (see Docherty in this issue), but the focus here is on Reading tasks only. Where possible, items were drawn from previously live tasks which were adapted for the revised format in order to provide a benchmark in terms of statistical performance.

The trial items were analysed using a Rasch model, which is a form of item response theory (IRT) model (see Elliott and Stevenson 2015:16–19). The analysis produced estimates of item difficulties in units called logits as well as an item discrimination statistic – point biserial correlation coefficient. The latter shows the extent to which an item distinguishes between strong and weak candidates.

FCE trial

The trial FCE candidature consisted of 317 candidates at 15 centres in eight countries (Elliott, Lim, Galaczi and Calver 2012). They were administered three shortened Reading tasks, consisting of 22 items in total: 1) a four-option multiple-choice task with one long text (6 items), b) a gapped-text task consisting of one text with missing sentences (6 items) and c) a multiple-matching task (10 items).

The trial showed that all Reading items but one (item 18) were within the acceptable range of item difficulties for FCE. As the items were previously used in a pretest, which is a process in which new tasks are trialled and the results analysed to determine item performance and difficulty using Rasch analysis (for detailed discussions of pretesting and use of Rasch analysis, see Corrigan and Crump 2015 and Elliott and Stevenson 2015 respectively), item trial difficulties were compared with their pretest difficulties. Evidently, most items performed similarly in both the trial and the pretest, with almost all variations being within the expected range due to sampling differences. Results showed that all statistical properties of the tasks such as difficulty and discrimination were within acceptable test construction ranges for the FCE Reading and Use of English test sections.

The mean difficulties of the trial tasks were close to the target difficulty of the exam, and the variations across tasks were smaller than in the pretest. Two tasks – multiple choice and multiple matching – were somewhat more difficult in the trial paper than in the pretest paper, while the gapped-text task was easier in the trial paper. The higher mean difficulty of the gapped-text task in the pretest was due to the exceptionally difficult item 12; were it not for this outlier, the difficulties of the task in the trial and the pretest would have been similar – only 1 scaled logit apart. The findings revealed that the shortened FCE Reading tasks perform well statistically: the items are within the acceptable difficulty range for FCE and they discriminate well between candidates.

CAE trial

The trial CAE candidature consisted of 137 candidates at 14 centres in 7 countries (Elliott et al 2012). They were administered three Reading tasks (alongside the Use of English ones) consisting of 23 items in total: 1) a four-option multiple-choice task with one long text (6 items), b) a gapped-text task consisting of one text with missing paragraphs (7 items in the trial task but 6 items in the revised task) and c) a multiple-matching task (10 items).

The trial showed that all trial Reading items but three (items 1, 3 and 7) were within the acceptable range of item difficulties for CAE. The trial item difficulties were compared with live and pretest difficulties of the same items. Most trial items performed similarly in the pretest compared to their previous live performance, with almost all variations being within the expected range due to sampling differences. As far as item discrimination is concerned, all items but one (at 0.21 point biserial) discriminated well between candidates, with the point biserial values ranging between 0.25 and 0.6. The findings revealed that the shortened CAE Reading tasks perform well statistically: a large majority of items are within the acceptable difficulty range for CAE and they discriminate well between candidates.

Refining Reading test constructs

As part of exam review, CAE and FCE were also investigated to ensure that their constructs reflected advances in research and that they are appropriate for the target candidature and intended contexts of use.

In the case of FCE, the aim was to make the test more suitable for Further Education students and those who would like to start working in an English-speaking environment (see Zeronis and Geranpayeh, and Howden and Mehta, this issue). This was achieved by including more work-related and adult education topics in the FCE Reading and Use of English paper. In the case of CAE, the aim was to make the test more suitable for those wishing to study at a university or for career advancement purposes (see Zeronis and Geranpayeh, and Howden and Mehta in this issue). This resulted in the identification of two areas for improvement in the CAE Reading test: a considerable overlap between two tasks and a gap in construct coverage.

The identified overlap between two CAE Reading tasks was between pre-revision Part 1 and Part 3 tasks. The two had the same task format (four-option multiple choice) and covered an identical range of testing focuses, such as 'detail, opinion, tone, purpose, main idea, implication, attitude, text organisation features (exemplification, comparison and reference)' (UCLES 2012:7); they also tested the same reading sub-skill – careful reading across sentences or paragraphs (Khalifa and Weir 2009:98). In view of this overlap, it was decided to exclude the Part 1 task from the revised test. This made way for a new, cross-text multiple-matching task (see Table 2 and the Appendix), which was designed to fill the identified gap in the pre-revision CAE Reading test construct.

The cross-test multiple-matching task assesses the ability to integrate information across several texts in order to critically evaluate the opinions expressed in them. None of the pre-revision CAE tasks tested this sub-skill, and even when a task required reading several texts (the multiple-matching task, see Table 2), focusing on each text in isolation was sufficient to respond to questions. Therefore, this was the cognitive operation which had not been explicitly or consistently elicited by Cambridge English Reading tests prior to the revised test. Integrating and evaluating information across texts is one of the key reading skills for following an academic course at university level and for working effectively in linguistically demanding professional contexts (Adler, Gujar, Harrison, O'Hara and Sellen 1998, Weir, Hawkey, Green, Ünaldi and Devi 2009). So, testing this aspect of reading ability is particularly important given that CAE is intended for prospective university students and for professionals interested in advancing their career. Crucially, the more representative a test is of real-life tasks, the more reliable inferences can be made on how well a test taker will be able to perform in a real-life environment.

The new CAE cross-text multiple-matching task

As part of a socio-cognitive framework (Weir 2005), representativeness of a reading test can be judged on its cognitive validity (Does it cover an adequately representative range of cognitive processes involved in reading?) and context validity (Are the contextual task features, such as language and topic, sufficiently representative of real-life texts?). To ensure cognitive and context validity of the new task, as well as its scoring validity (How well does the task perform statistically?), quantitative and qualitative analyses were carried out within a mixed methods approach (for more detail on the approach and the trial, see Elliott and Lim forthcoming 2016). The new task was developed in three phases.

Phase 1

In the initial phase, text and task features were considered. In view of considerable discipline-specific variations in academic (and professional) discourse and genres, it was decided that more general texts which share the higher-level features of academic texts should be selected for this task. The intention was to avoid the problem of the inaccessibility of discipline-specific language to a wide range of test takers, and to allow texts to contain a general argumentative structure of an academic text. As far as task types are concerned, previous research has shown that certain task formats, such as multiple-choice questions and gap filling, typically elicit lower-level reading, directing test takers to focus predominantly on discrete points of information and comprehension at word, phrase and sentence level (Cohen 1984, Nevo 1989, Rupp, Ferne and Choi 2006, Weir, Hawkey, Green and Devi 2009). However, a multiple-matching task where texts function as answer options would broaden task focus and elicit reading across texts. Therefore, this task type was chosen as the most appropriate. Since there was already one multiple-matching task in the pre-revision CAE, two options were trialled: 1) adapting the existing multiple-matching task by adding two items which focus on inter-textual reading comprehension, while retaining the existing testing focus for other items; and, 2) developing a new multiple-matching task focusing only on inter-textual reading comprehension.

Two versions of the adapted task and two versions of the new cross-text task were created, with each version consisting of four evaluative texts on the same theme (for example, reviews of a book). The items were written in order to test whether the candidate can identify agreement or disagreement between authors, with the texts themselves forming the four answer options. There were two sets of trial items: those which require identifying an opinion expressed in one of the texts and then identifying which other text shares or contradicts this opinion, and those which require identifying which text differs from the others in terms of an expressed opinion. In both cases, candidates must select one text only; items only provide information on the subject of the opinion but not the opinion itself, which the candidate must identify.

The texts were then evaluated, using expert judgement, for appropriateness of text purpose, style/register, as well as functional, lexical and grammatical features. The tasks were also evaluated to determine if they elicit appropriate types of reading. The texts in the new tasks were found more appropriate in terms of contextual features than those in the adapted tasks, and the new tasks also elicited the intended reading type:

> The texts for [the new task] are of a nature consistent with academic texts in terms not only of vocabulary, structures and lexical bundles but also in their expository/argumentative overall text purposes and their detached tone and formal style. This contrasts with the descriptive/narrative, informal texts of a personal nature in the tasks for [the adapted task], which do not contain the lexical or grammatical complexity of their counterparts in [the new task].
>
> Critically, and as a consequence of the features described above, it is necessary to read across stretches of text in order to locate the answers to the items in New Tasks 1 and 2, whereas the information required to respond to the items in Adapted Tasks 1 and 2 is found locally within individual sentences (and across no more than three sentences), and is more explicitly stated. This means that considerably more higher-level processing is engaged by [the new task], whereas [the adapted task] may only require processing up to the level of individual propositions (Elliott, Vidaković and Corrigan 2013:2).

Phase 2

Following the qualitative (content) analysis, two versions of the new task and two versions of the adapted task were trialled along with the remainder of Reading and Use of English tasks. The Rasch analysis and standard classical analysis were based on a sample of 150 CAE candidates. The sample was adequately balanced in terms of the first language background to avoid language-specific bias (i.e. no more than one third of candidates from a given language group). Item difficulty, facility and discrimination were determined for each new and adapted cross-text item and task. The task which was found to perform acceptably well in the statistical sense, and to satisfy context and cognitive validity criteria, was one of the two new tasks. It was chosen as the most viable cross-text task and task specifications for item writers were refined by drawing on both quantitative and qualitative findings. Based on the revised task specifications, 27 cross-text tasks were produced for the next round of trialling.

Phase 3

In the final phase of task development, the 27 cross-text tasks were analysed quantitatively to determine the extent to which the texts exhibit similar properties to the texts encountered in the Higher Education context.

Following Green, Ünaldi and Weir (2010) and Green, Weir, Chan, Taylor, Field, Nakatsuhara and Bax (2012), the texts were then quantitatively analysed using Coh-Metrix 3.0 (McNamara, Louwerse, Cai and Graesser 2012), which is an automated, web-hosted 'computational tool that provides a wide range of language and discourse measures . . . that [users] can use to obtain information about their texts on numerous levels of language' (McNamara, Graesser, McCarthy and Cai 2014:1) to investigate lexical and syntactic complexity and text coherence and cohesion. VocabProfile (Cobb 2003) was used to provide additional lexical measures. The texts totalled 16,009 words and were compared with an undergraduate mini-corpus from the Centre for Research in English Language Learning and Assessment (CRELLA) consisting of 42 extracts from 14 texts, totalling 18,484 words. The analysis detailed lexical, syntactic and coherence/cohesion properties of the texts, allowing for a comparison of those features which exhibited statistically significant differences between the CAE texts and the undergraduate texts and which may be used as indicators of the relative reading difficulty of the texts.

The results of the analysis indicated that as far as lexical characteristics are concerned, CAE texts are similar to those in the analysed undergraduate corpus in terms of word length, the proportion of academic words and infrequent (more sophisticated) words. There is no indication that CAE texts are less lexically challenging. Moreover, the type-token ratio shows that CAE texts are lexically more diverse than those in the undergraduate corpus.

A syntactic analysis showed that adjacent sentences in CAE texts in the new task are less thematically related than in the undergraduate corpus, which indicates a source of increased

difficulty in the CAE texts. However, a lower mean number of modifiers per noun phrase in CAE should facilitate reading and make the text easier. The results suggest little difference in syntactic complexity between the two sets of texts, given that the majority of indices are not statistically significantly different and that the two which are differ in opposite directions by moderate amounts (the differences in both cases are no more than the standard deviation of the undergraduate text scores).

An analysis of textual coherence and cohesion showed statistically significant differences between the two sets of texts. The undergraduate texts returned higher scores, which suggests features which should facilitate reading of the undergraduate texts compared to the CAE texts. This is, perhaps, not surprising given that the CAE texts each comprise four separate mini-texts, so coherence across texts (even though they are thematically related) should be expected to be lower than within a single text from the undergraduate corpus.

The conclusion arising from the findings is that CAE and undergraduate texts are similar in terms of lexical, syntactic and textual complexity, with no evidence that they have lower reading difficulty levels. This represents strong evidence to support the context validity of the new task in relation to the Higher Education context. A full discussion of the Coh-Metrix analysis, including detailed analysis of the data, can be found in Elliott and Lim (forthcoming 2016).

Conclusion

The revision of FCE and CAE Reading papers resulted in shortening the tasks and merging Reading and Use of English. The shortening of the papers has not had a negative impact on item and task performance, as they were shown to remain at an appropriate level of difficulty and to discriminate well between candidates. The key change was including a new, inter-textual Reading task in CAE. This considerably broadened the exam construct to make it more suitable for the target candidature and more representative of the reading skills required in a Higher Education or professional environment.

A key consideration for CAE was bolstering its validity for use as a Higher Education entry requirement, without straying into the realms of English for Specific Purposes (ESP) testing; in this sense, CAE represents evidence that a candidate possesses the requisite level of language to cope with university study without presuming that they are familiar with the specific domain requirements. Consideration was given to texts featuring lexical and syntactic characteristics similar to those of undergraduate texts, but whose treatment of a topic did not presume specialist in-depth knowledge.

The new format provides a clear progression from FCE to CPE in terms of levels and testing focuses of the tasks. In the multiple-choice task all three exams have a similar testing focus, and difficulty stems from increasing text complexity and level of abstraction. In the gapped text task, FCE focuses testing on more localised reading at sentence level, progressing to gapped paragraphs in CAE and CPE where the testing focus encourages global reading for features of textual coherence and cohesion. The multiple-matching task tests reading for specific information, detail, opinion and attitude at all three levels, but again there is a progression from B2 to C levels in terms of text complexity and increasing text length.

In order to inform the revision and reach a well-rounded insight into the appropriateness of the new tests, all changes were trialled using quantitative and qualitative methods as part of a mixed methods approach. Moreover, different aspects of the revised tests were investigated using a socio-cognitive framework for test development and validation (Weir 2005). This evidences a rigorous and multi-faceted approach to test validation and revision. As test validation is always an ongoing process, we are planning to carry out further research on the revised tests. In particular, the intention is to investigate the cognitive validity of the CAE cross-text task by employing techniques such as verbal think-aloud protocols, retrospective questionnaires and eye-tracking, in order to determine how test takers engage with the new task.

References

Adler, A, Gujar, A, Harrison, B L, O'Hara, K and Sellen, A (1998) A diary study of work-related reading: Design implications for digital reading services, in *Proceedings of the SICCHI Conference on Human Factors in Computing Systems*, Boston: ACM Press/Addison Wesley Publishing Co, 241–248.

Cobb, T (2003) *Web VocabProfile*, available online: www.lextutor.ca/vp

Cohen, A (1984) On taking language tests: what the students report, *Language Testing* 1 (1), 70–82.

Corrigan, M and Crump, P (2015) Item analysis, *Research Notes* 59, 4–9.

Elliott, M and Lim, G S (forthcoming 2016) The development of a new reading task: A mixed methods approach, in Moeller, A J, Creswell, J W and Saville, N (Eds) *Second Language Assessment and Mixed Methods Research*, Studies in Language Testing volume 43, Cambridge: UCLES/Cambridge University Press.

Elliott, M and Stevenson, L (2015) Grading and test equating, *Research Notes* 59, 14–20.

Elliott, M, Vidaković, I and Corrigan, M (2013) *CAE Reading and Use of English Trial 5 Report*, Cambridge: Cambridge English internal report.

Elliott, M, Lim, G S, Galaczi, E D and Calver, L (2012) *FCE and CAE construct validation study (Part 2)*, Cambridge: Cambridge English internal report.

Green, A, Ünaldi, A and Weir, C J (2010) Empiricism versus connoisseurship: establishing the appropriacy of texts for testing reading for academic purposes, *Language Testing* 27 (3), 1–21.

Green, A, Weir, C J, Chan, S, Taylor, L, Field, J, Nakatsuhara, F and Bax, S (2012) *Textual features of CAE reading texts: CAE texts compared with reading texts from FCE, CPE, IELTS and with essential undergraduate textbooks*, final project report, University of Bedfordshire.

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.

McNamara, D S, Graesser, A C, McCarthy, P M and Cai, Z (2014) *Automated Evaluation of Test and Discourse with Coh-Metrix*, New York: Cambridge University Press.

McNamara, D S, Louwerse, M M, Cai, Z and Graesser, A (2012) *Coh-Metrix version 3.0*, available online: cohmetrix.memphis.edu

Nevo, N (1989) Test-taking strategies on a MC test of reading comprehension, *Language Testing* 6 (2), 199–217.

Rupp, A A, Ferne, T and Choi, H (2006) How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective, *Language Testing* 23 (4), 441–474.

UCLES (2012) *Cambridge English: Advanced: Handbook for Teachers*, Cambridge: UCLES.

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

Weir, C J, Hawkey, R, Green, A and Devi, S (2009) The cognitive processes underlying the academic reading construct as measured by IELTS, *IELTS Research Reports Volume 9*, available online: www.ielts.org/pdf/Vol9_Contents_Page.pdf

Weir, C J, Hawkey, R, Green, A, Ünaldi, A and Devi, S (2009) The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university, *IELTS Research Reports Volume 9*, available online: www.ielts.org/pdf/Vol9_Contents_Page.pdf

Williams, E and Moran, C (1989) Reading in a foreign language at intermediate and advanced levels with particular reference to English, *Language Teaching* 22 (4), 217–228.

Zeronis, R and Elliott, M (2013) Development and construct of revised Cambridge English: Proficiency, *Research Notes* 51, 22–31.

Appendix

The new CAE Reading cross-text multiple-matching task

Part 6

You are going to read four extracts from articles in which academics discuss the contribution the arts (music, painting, literature, etc.) make to society. For questions **1 – 4**, choose from the academics **A – D**. The academics may be chosen more than once.

Mark your answers **on the separate answer sheet**.

The Contribution of the Arts to Society

A Lana Esslett

The arts matter because they link society to its past, a people to its inherited store of ideas, images and words; yet the arts challenge those links in order to find ways of exploring new paths and ventures. I remain sceptical of claims that humanity's love of the arts somehow reflects some inherent inclination, fundamental to the human race. However, exposure to and study of the arts does strengthen the individual and fosters independence in the face of the pressures of the mass, the characterless, the undifferentiated. And just as the sciences support the technology sector, the arts stimulate the growth of a creative sector in the economy. Yet, true as this is, it seems to me to miss the point. The value of the arts is not to be defined as if they were just another economic lever to be pulled. The arts can fail every measurable objective set by economists, yet retain their intrinsic value to humanity.

B Seth North

Without a doubt, the arts are at the very centre of society and innate in every human being. My personal, though admittedly controversial, belief is that the benefits to both individuals and society of studying science and technology, in preference to arts subjects, are vastly overrated. It must be said, however, that despite the claims frequently made for the civilising power of the arts, to my mind the obvious question arises: Why are people who are undeniably intolerant and selfish still capable of enjoying poetry or appreciating good music? For me, a more convincing argument in favour of the arts concerns their economic value. Needless to say, discovering how much the arts contribute to society in this way involves gathering a vast amount of data and then evaluating how much this affects the economy as a whole, which is by no means straightforward.

C Heather Charlton

It goes without saying that end-products of artistic endeavour can be seen as commodities which can be traded and exported, and so add to the wealth of individuals and societies. While this is undeniably a substantial argument in favour of the arts, we should not lose sight of those equally fundamental contributions they make which cannot be easily translated into measurable social and economic value. Anthropologists have never found a society without the arts in one form or another. They have concluded, and I have no reason not to concur, that humanity has a natural aesthetic sense which is biologically determined. It is by the exercise of this sense that we create works of art which symbolise social meanings and over time pass on values which help to give the community its sense of identity, and which contribute enormously to its self-respect.

D Mike Konecki

Studies have long linked involvement in the arts to increased complexity of thinking and greater self-esteem. Nobody today, and rightly so in my view, would challenge the huge importance of maths and science as core disciplines. Nevertheless, sole emphasis on these in preference to the arts fails to promote the integrated left/right-brain thinking in students that the future increasingly demands, and on which a healthy economy now undoubtedly relies. More significantly, I believe that in an age of dull uniformity, the arts enable each person to express his or her uniqueness. Yet while these benefits are enormous, we participate in the arts because of an instinctive human need for inspiration, delight, joy. The arts are an enlightening and humanising force, encouraging us to come together with people whose beliefs and lives may be different from our own. They encourage us to listen and to celebrate what connects us, instead of retreating behind what drives us apart.

Which academic

has a different view from North regarding the effect of the arts on behaviour towards others? **1** ☐

has a different view from Konecki on the value of studying the arts compared to other academic subjects? **2** ☐

expresses a different opinion to the others on whether the human species has a genetic predisposition towards the arts? **3** ☐

expresses a similar view to Esslett on how the arts relate to demands to conform? **4** ☐

Revising the Use of English component in FCE and CAE

COREEN DOCHERTY RESEARCH AND THOUGHT LEADERSHIP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Introduction

As part of the regular exam review process for the *First Certificate in English* (FCE) and *Certificate in Advanced English* (CAE) exams (see Zeronis and Geranpayeh, in this issue), each individual paper, and both exams as a whole, were evaluated in terms of their fitness for purpose and the extent to which they are in line with current knowledge about measuring language proficiency. This article focuses on the modifications made to the FCE and CAE Use of English components, which are designed to measure grammatical and lexical knowledge. A summary of the main changes to these papers includes the following:

1. The Reading and Use of English components were merged into one paper, but continue to be reported separately.
2. Text-based Use of English tasks were shortened and the number of items reduced.
3. The gapped-sentences task in CAE was removed.
4. The multiple-choice cloze task now contributes to the Reading component score rather than the Use of English one.

As the Reading and Use of English components have been combined into one paper, this article complements the one by Vidaković, Elliott and Sladden in this issue, which describes the changes to the Reading papers.

Measuring language knowledge

A key consideration of this review, and all Cambridge English exam reviews, is the model of language proficiency underlying these exams, which is based on the notion that communicative language ability can be divided into different sub-skills and abilities (Geranpayeh 2007). That is, although an overall language ability exists, language skills (i.e. reading, writing, speaking and listening) and language knowledge or systems (i.e. grammar and vocabulary) can develop differently in each individual and can be measured as separate aspects of language ability. As such, a feature of many Cambridge English exams is that they are comprised of different components or papers, which allow candidates to demonstrate their mastery in each. This enables Cambridge English to report not only on a candidate's overall proficiency but also provide more granular information for each skill, which can be used to support further learning by identifying the skills which are strong or weak. This approach to exam design is aimed at supporting positive impact, which is especially important for general and academically oriented English exams such as FCE and CAE, which are frequently used in the school sector. That is, treating language ability as componential encourages an explicit focus on all skills and systems equally in the classroom (Ashton, Salamoura and Diaz 2012, Chambers, Elliott and Jianguo 2012, Docherty, Casacuberta, Pazos and Canosa 2014).

Although Cambridge English considers language ability divisible into skills and systems, the components included in a particular exam may differ depending on the test purpose and/or Common European Framework of Reference (CEFR, Council of Europe 2001) level of the exam. Generally, the majority of exams test reading, listening, writing, speaking and language knowledge to some extent, with some skills being prioritised or not included based on the target language use context. For example, a test for call centre employees may prioritise listening and speaking over reading and writing because of the nature of the job; therefore, the test may have separate Listening and Speaking components but a combined Reading and Writing paper with more weight given to the former skills than the latter when scores are combined to determine an overall grade. Similarly, the CEFR level of the exam may also influence which components are included or tested separately. At lower CEFR levels (B1 and below), for instance, there is a clear overlap between language knowledge and reading and writing because learners have such a limited language repertoire. That is, below B1, reading and writing tends to involve lower-level cognitive processing such as lexical and grammatical recognition and retrieval, which makes it difficult to distinguish between these language skills and the underlying lexical and grammatical abilities. While language knowledge is an enabling skill which underlies all four skills (it would not be possible to engage in communication without some lexical and grammatical knowledge, and as Weir and Porter (1994:8) argue 'it does seem improbable that students would be able to work out the main ideas of a text without some baseline competence with the microlinguistic skills) the particularly strong correlation with reading and writing (Nation 2007, Purpura 2004, Read 2000) is reflected in the Cambridge English test constructs. Thus, Cambridge English exams at B1 and below generally test language knowledge explicitly as part of Reading and/or Writing papers partly because lexical and grammatical knowledge sit comfortably alongside these two skills, and also to reduce test length and potentially negative test impact. A test designed to measure language knowledge separately from the other skills at the lower CEFR levels would need to be unnecessarily long in order to gather enough information to reliably report scores for these different components of language ability. At the lower CEFR levels it may not be worth the drawback of having candidates sit a longer test. At this level, a longer test may have negative consequences on performance or on the candidates' attitudes towards the testing experience. A further example of this is that the lower level reading processes in the Khalifa and Weir's (2009) reading model (see Vidaković, Elliott and Sladden, in this issue) rely on word recognition and lexical parsing which are related to language knowledge, whereas in writing, lexical resource and linguistic patterns are the main features at the lower CEFR levels.

However, from CEFR Level B2 (i.e. FCE) upwards, language

knowledge is tested as a separate component and reported as one aspect of overall language proficiency alongside reading, listening, speaking and writing. This is because B2 level learners have a wider range of skills and abilities, including the ability to engage in higher-order reading processes (Khalifa and Weir 2009, Vidaković, Elliott and Sladden, this issue) and in more complex writing activities such as the transformation of knowledge. It becomes easier to distinguish between the lower-level processing associated with language knowledge and higher-order processes associated with skilled readers or writers at B2 level and above. Additionally, a separate Use of English paper enables Cambridge English to more accurately and reliably make inferences about a learner's overall proficiency than when it is embedded in other skills. The separate Use of English paper gives candidates a range of tasks, allowing them to demonstrate their language knowledge and their ability to use this knowledge productively, which may not be possible within skill-based papers because of topic or contextual constraints. There is also less of a need to explicitly test language knowledge in reading and listening papers if it is tested separately, thus freeing up valuable space that can be used for testing higher-order cognitive processes.

A lexico-grammatical approach

A Use of English component was first introduced into a Cambridge English exam in the 1950s and the testing focus of this component has evolved in accordance with changes to teaching and testing over time (Weir 2013). In line with a communicative approach to teaching and assessing, the testing focus of FCE and CAE Use of English papers is on 'lexico-grammatical competence which includes components of meaning, word formation, collocation, lexical relationships, lexical cohesion, modality, complementation, phrase structuring, clause combining and grammatical cohesion' (Hawkey 2009:82). This focus goes beyond simply knowing about these aspects of language but includes the ability to apply this knowledge, thus linking form, meaning and use together.

Cambridge English takes a lexico-grammatical approach in order to emphasise the relationship between grammar and vocabulary. It has been argued that trying to disentangle grammar and vocabulary into separate components is both challenging and perhaps not justified when teaching and testing (Celce-Murcia and Larsen-Freeman 1999, Halliday 1985, Shiotsu and Weir 2007). Recent empirical support for this view comes from research conducted as part of the English Profile Programme, which is developing Reference Level Descriptions for English that aim to identify criterial features for each CEFR level. The English Vocabulary Profile and English Grammar Profile are two such projects. Using learner productive output from the Cambridge Learner Corpus (described in Harrison 2015) and numerous other resources (see Capel 2010, 2012), these projects have made great strides in identifying how learners' knowledge of structure and lexis develops. Hawkins and Filipović (2012) found that there is a strong relationship between vocabulary and grammar for learners in that certain vocabulary items trigger particular grammatical features. Learners' grammatical development is 'often relative to [their] vocabulary development (Harrison 2015:34). Learners do not simply learn a grammatical form and all aspects of it at a particular CEFR level, but rather a form is learned alongside a limited number of lexical items at first; then, as their language develops, the range of lexical items that they use alongside the grammatical feature expands, while at the same time they develop their ability to recognise the range of meanings a grammatical structure can have (Harrison 2015).

Summary of the Use of English component: Pre- and post-revision

Taking into consideration the conceptualisation of language knowledge described above, the components were updated. Tables 1 and 2 provide an overview of the changes to the FCE and CAE Use of English components respectively.

Table 1: Comparison of pre- and post-revision FCE Use of English component

Use of English tasks	Task focus	Pre-revision	Post-revision
Multiple-choice cloze	Unchanged: Lexical/ lexico-grammatical	12 items 180–200 words	8 items 150–160 words
Open cloze	Unchanged: Grammatical/ lexico-grammatical	12 items 180–200 words	8 items 150–160 words
Word formation	Unchanged: Lexical/ lexico-grammatical	10 items 180–200 words	8 items 150–160 words
Key word transformations	Unchanged: Lexical and grammatical	8 items c. 20 words each	6 items c. 20 words each
Total items/ words		42 items 700–760 words	30 items 570–600 words

Table 2: Comparison of pre- and post-revision CAE Use of English component

Use of English tasks	Task focus	Pre-revision	Post-revision
Multiple-choice cloze	Unchanged: Lexical/lexico-grammatical	12 items 210–230 words	8 items 150–170 words
Open cloze	Unchanged: Grammatical/ lexico-grammatical	15 items 220–250 words	8 items 150–170 words
Word formation	Unchanged: Lexical/lexico-grammatical	10 items 190–210 words	8 items 150–170 words
Gapped sentences*	Lexical	5 items c. 70 words each	Removed
Key word transformations	Unchanged: Lexical and grammatical	8 items c. 25 words each	6 items c 25 words each
Total items/ words		50 items 1,170–1,240 words	30 items 600–660 words

*These tasks were removed

As can be seen in Tables 1 and 2, the text-based task types (i.e. multiple-choice cloze, open cloze and word formation) and the testing focus of these tasks remain the same for both papers pre- and post-revision; however, the length of passages and the number of items have been reduced. Similarly, both papers have retained key word transformations but there are

fewer items in the post-revision versions of the papers. Finally, the gapped-sentences task in CAE has been removed. These changes bring the two components more clearly in line with each other while still providing a clear progression from FCE to CAE in terms of the word length of input texts. More detailed information on the testing focus, input text type, response format and scoring procedures can be seen in Table 3.

Table 3: Structure of post-revision FCE and CAE Use of English exams

Part	Task type	Testing focus	Input text type	Response format	Score
1	Multiple-choice cloze	Vocabulary: recognition of precise meaning, collocations, fixed phrases, phrasal verbs, etc. An element of grammatical knowledge may also be involved.	Short passage	Selected (4 options)	0, 1
2	Open cloze	Grammatical knowledge and use: grammatical structure and also some features of textual cohesion. Spelling.	Short passage	Productive	0, 1
3	Word formation	Lexical and morphosyntactic knowledge and use: knowledge of word formation, including affixation of prefixes and suffixes, internal changes and compounding. Spelling.	Short passage	Productive	0, 1
4	Key word transform-ations	Lexical and grammatical knowledge, meaning and use: greater emphasis is given to testing structure. Spelling.	Discrete sentences	Productive	0, 1, 2

The paper is designed in an attempt to promote a communicative approach to language testing. The multiple-choice cloze is a rational cloze, which means a specific class of words are removed from the text such as pronouns, lexical items, prepositions, etc. rather than deleting words at a regular interval, e.g. every 10 words, which will result in a variety of word classes being removed and not necessarily a coherent set. The main aim of the multiple-choice cloze is to assess learners' vocabulary knowledge, including their understanding of collocation, fixed phrases and phrasal verbs. This may involve an inherent grammatical element, because part of knowing how a word is used is recognising its grammatical constraints. For example, learners may need to recognise the correct complementation of an item (e.g. which preposition or which verb form follows). In this task, learners are given four options to choose from and receive one mark for a correct choice. The open cloze, which is also a rational cloze, tends to focus on grammatical knowledge but also includes lexical items as part of fixed phrases, and items can also tap into textual cohesion. Candidates are not given options but rather must produce one word to fill the gap, which must be spelled correctly in order to receive one mark.

The word formation task tests morphosyntactic knowledge, which includes affixation of prefixes and suffixes, internal changes and compounding. This task also requires learners to produce the new form of the word, spelled correctly, in order to receive one mark. Finally, the last Use of English task, key word transformations, requires learners to manipulate structure and lexis in order to produce sentences similar in meaning to the input provided. The testing focus is on both lexis and structure in that the key word will often trigger a particular grammatical form. This task provides information on learners' lexico-grammatical range and is the only discrete task on the paper. Although lexico-grammatical range and accuracy are also tested in the Writing and Speaking papers, the topic and each learner's own choices may limit the grammatical and lexical range they produce. This task allows learners to demonstrate their full linguistic repertoire. Unlike the other Use of English tasks where each correct answer receives one mark, candidates can receive up to two marks (i.e. 0, 1 or 2) on each key word transformation item, allowing partially correct answers to be recognised.

As can be seen in this summary of tasks, the Use of English component emphasises productive tasks over selected-response formats so that candidates are required to demonstrate not only their knowledge of language but also the ability to use this knowledge productively, which is a more cognitively demanding task. This is an important feature for the testing of grammar and vocabulary as a learner's language knowledge and the level of control they have over this knowledge may be quite different. Additionally, in order to improve the authenticity of assessment tasks and make them more communicative, text-based task types which test knowledge in context are prioritised over discrete tasks. Text-based tasks where learners may need to complete a gap by selecting or providing the appropriate word embeds the language knowledge construct within the larger construct of reading. The cognitive processes activated in these tasks: recognising words, lexical parsing, reading at the phrase level, sentence level and on occasion beyond the sentence level, are the lower-level reading processes described in the Khalifa and Weir (2009) reading model (see Vidaković, Elliott and Sladden, this issue). This reinforces the close link between Reading and Use of English tasks and mimics instructional tasks, as Purpura (2004) points out, which adds an element of authenticity. Both these features, the inclusion of productive items and testing language knowledge in context, should encourage positive washback in the classroom because of the focus on language use rather than solely on language form.

Although the task types and the overarching testing focuses are the same for both exams, they differ in the range and depth of lexical and grammatical knowledge candidates are expected to have. The linguistic complexity of the items increases from B2 to C1 level in that it is expected that learners not only have a larger repertoire of language knowledge at their disposal but that they have more control over this repertoire so that they are able to use it flexibly. For example, distinguishing between a correct and incorrect option in the CAE exam is based on a more advanced understanding of language form, meaning and use than in FCE. That is, items at the C1 level are intended to tap into a learner's understanding that particular grammatical forms can have more than one function. Additionally, the basis for

identifying the correct answer for gap-fill tasks (Parts 1–3) in CAE may rely on processing information in preceding or proceeding sentences – a higher-order cognitive process – whereas in FCE the cognitive processing activated is more likely to be restricted to the level of the phrase and sentence (see Vidaković, Elliott and Sladden, this issue).

The changes summarised above were based on a number of considerations including the desire to shorten the length of the exam by eliminating duplication in testing focus, ensure the exam is suitable for work, study and general purposes and reflect current research on assessing language knowledge. A number of activities, therefore, were undertaken as part of the revision process such as examining the statistical performance of each task and the paper as a whole, reviewing the current test specifications and analysing recent tests to examine the constructs covered in the paper, and considering alternative task types. During this process, the gapped-sentences task in CAE, which focuses on lexical knowledge, was selected for removal primarily because this testing focus is captured elsewhere on the paper (i.e. on the multiple-choice cloze, word formation and the key word transformations tasks). Additionally, this task consisted of discrete sentences and there was a preference to prioritise the text-based tasks, which are more conducive to testing features beyond the sentence and may discourage instructional practices which focus on memorisation or rote learning of vocabulary. Finally, using a technique called Structural Equation Modelling, which is a statistical method for testing conceptual or theoretical models such as the componentiality of language proficiency (i.e. is language ability a unitary concept or is it divisible into separate skills and systems?), Geranpayeh and Somers (2006) found that the gapped-sentences task in the pre-revision CAE exam was less consistent with the rest of the Use of English tasks and had a stronger relationship with the Reading paper. This is not surprising considering this task most clearly focuses on breadth and depth of vocabulary knowledge, which research suggests is highly correlated to reading ability (Nation 2001, Read 2000). For these reasons, it was felt that the removal of this task would not lead to construct underrepresentation. The remaining tasks were deemed to be sufficiently different in testing focus to be retained.

The next revision activity centred on the text-based tasks (Parts 1–3) and whether the text length and number of items could be reduced and still maintain their construct coverage and perform statistically at the level. The qualitative analysis of pre-revision tasks indicated that there was a tendency for multiple items within the same task and across tasks to have a similar testing focus such as verb + noun collocation. As such it was possible to have fewer items per task and still maintain the overall range of features tested in each task and the paper as a whole. Once it had been determined that the tasks could be shortened, the ratio of items-to-text needed to be considered as this can affect item difficulty (Abraham and Chapelle 1992, Alderson 2000) and the type of cognitive processing activated. For example, in CAE items have always been included in the text-based tasks which are expected to activate higher-level processing such as understanding meaning across sentences. Learners at C1 level are able to read complex texts and use higher-level reading processes so it is expected that they are able to go beyond the immediate surroundings of a gap to answer an item by identifying the relevant information in neighbouring sentences. It was seen as important that this feature, which distinguishes this exam from lower-level exams, would not be eliminated as a result of shortening the texts. Different text lengths and number of items were trialled (discussed in the next section) until the optimal text-to-item ratio was identified and as part of this process, qualitative analyses of tasks was undertaken to monitor the cognitive processes likely to be activated.

The changes made to the Use of English papers reduced the amount of time needed to complete these papers while still providing enough information to report a separate Use of English score for candidates. Text-based tasks were retained and discrete tasks were reduced to encourage an emphasis on language knowledge in context (e.g. a focus on meaning and use) rather than an overemphasis on form. It is hoped that this focus on meaning and use will have a positive washback on classroom teaching.

Trialling

During the review process, multiple trials were organised with candidates who were preparing for one of the exams (see Vidaković, Elliott and Sladden, this issue, for more details about the trials). The early trials focused on 'proof of concept'. Although there were no new tasks included in the Use of English component, it was necessary to determine whether the changes to text length and number of items for Parts 1–3 (i.e. multiple-choice cloze, open cloze and word-formation tasks) would affect how they functioned. Qualitative analysis of tasks was undertaken to investigate the cognitive processes which would likely be activated while quantitative analysis was used to determine whether items and tasks performed statistically as expected in terms of item difficulty and discrimination (i.e. the ability of an item to distinguish between stronger and weaker candidates). As mentioned previously, changing the item-to-text ratio in gap-fill tasks can affect the difficulty of items and with a higher number of words per gap in the revised tasks, it was possible that they would be too easy for their respective CEFR levels. Trials demonstrated that items were not easier and performed in a similar way to the pre-revision tasks. This may be the case because the shorter, more concise passages do not give much scope for extrapolations or explanations to support the development of an argument.

Once it had been determined that the format change had not affected the testing focus and the difficulty of the test, further trials were conducted to monitor the statistical performance of the tasks, determine the appropriate length of time to give candidates to complete the test and determine the optimal order of tasks within the test. For example, there was some concern that if the Use of English tasks were first in the paper, candidates may spend too much of their time on them and then run out of time before finishing the Reading section. As such, trial observers were asked to pay attention to how candidates worked their way through the exam: did they start with the Reading component and then go back to the Use of English component or did they follow the order of the test and did this choice result in the candidate running out of time? In addition, for each trial candidate and teacher, perceptions of the test in terms of difficulty, timing and

appropriacy of task format and content for the CEFR level were sought.

The results of trialling indicated that the shorter text-based tasks performed in a similar manner to the longer passages and they were not perceived to be any more or less difficult than the longer passages. The mean overall difficulty of the component was very close to the original pretest mean of the items in their original, longer form and almost all items performed similarly in pretests and trials, with almost all variations within the expected range due to sampling differences. It was also determined that the Use of English tasks were better positioned first in the Reading and Use of English paper because there was no indication that this negatively affected exam performance and it allowed these tasks to act as a bridge to the following tasks as they tend to activate lower-level cognitive processing within the reading model described in Khalifa and Weir (2009). Most critically, teachers did not raise concerns over the merging of the two components and they also did not perceive the new shorter tasks as lowering the level of the exam, which was a concern raised during the stakeholder consultation phase (see Howden and Mehta, this issue). However, one outcome of the trials was that the multiple-choice cloze task was moved to the Reading paper. Different permutations of aggregating scores were considered with the multiple-choice cloze included as part of the Use of English score and as part of the Reading score (see Table 4). Tables 5 and 6 show the internal comparisons of the Alpha and Standard Error of Measurement (SEM) figures for the two permutations for each exam (Elliott, Lim, Galaczi and Calver 2012). Both the Alpha and SEM are reliability indicators with the Alpha value describing the internal consistency of the items and the SEM value indicating the extent to which a candidate's score may fluctuate if they were to take the test again (Somers 2015). Typically an Alpha value above 0.80 is considered acceptable for a component of a test and a low SEM figure is valued. Tables 5 and 6 indicate that Permutation B produced more equal Alphas and SEMs for both components, whereas Permutation A produced a higher Use of English Alpha (and lower SEM) at the expense of a lower Reading Alpha (and higher SEM).

Although both permutations produce higher SEMs for Use of English, this is a natural consequence of the reduction in items (and information points) in the component, and is not necessarily a cause for concern; the pre-revision Use of English formats contain considerably more information points than their respective Reading or Listening counterparts, and the increased SEM in the post-revision component only serves to bring it approximately in line with the figures for those papers (Elliott et al 2012).

As a result, scores on the multiple-choice cloze task now contribute to the Reading score, which is an essential change to the test construct as a result of merging the Reading and Use of English components. This close link between lexical knowledge and reading comprehension is further supported when investigating the constructs of test components using Structural Equation Modelling techniques. In those studies, the multiple-choice cloze task with a focus on vocabulary was shown to be associated strongly with Reading tasks in both exams (Elliott, Docherty and Benjamin 2015, Geranpayeh and Somers 2006, Malarkey and Somers 2012). In fact, this task has appeared variously in both Reading and Use of English papers historically, which highlights its flexibility. The multiple-choice cloze task continues to be grouped, however, with the Use of English tasks as it appears first in the paper. This decision is based on task type, format and perceived difficulty. It is a task that is often associated with the Use of English paper and it is the only selected-response Use of English-type task, which is a format that can be perceived as easier because candidates do not need to produce language. Therefore, it is viewed as a task that can ease candidates into the paper.

Table 4: FCE and CAE Reading/Use of English score reporting permutations

	Task 1 (MC* cloze)	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7
Permutation A	UoE	UoE	UoE	UoE	Reading	Reading	Reading
Permutation B	Reading	UoE	UoE	UoE	Reading	Reading	Reading

*Multiple-choice

Table 5: FCE Reading and Use of English summary statistics for candidate performance in Trial 2 versus 2011 live administrations (scaled to 40 marks maximum)

		Permutation A	Permutation B	2011 live range
Reading	Alpha	0.86	0.87	0.81–0.88
	SEM	3.68	3.24	2.88–3.33
Use of English	Alpha	0.89	0.87	0.83–0.89
	SEM	2.91	3.32	2.33–2.63

Table 6: CAE Reading and Use of English summary statistics for candidate performance in Trial 2 versus 2011 live administrations (scaled to 40 marks maximum)

		Permutation A	Permutation B	2011 live range
Reading	Alpha	0.85	0.87	0.76–0.85
	SEM	3.46	3.00	2.74–3.11
Use of English	Alpha	0.85	0.84	0.79–0.87
	SEM	2.81	3.32	2.30–2.43

Since the paper has gone live, reliability and SEM values for the Use of English components for both papers continue to be in line with historical values. Table 7 shows the summary statistics for one of the largest sessions in 2015 for FCE and CAE.

Table 7: Reliability figures for FCE and CAE Use of English components

	FCE	CAE
Alpha	0.79	0.79
SEM	2.43	2.49

Conclusion

The modifications to FCE and CAE have resulted in a shorter Use of English element and a combined Reading and Use of English paper. The main changes to the Use of English component include shorter texts and fewer items, and the removal of the gapped-sentences task in CAE. These changes do not appear to have reduced construct coverage and it is still possible to distinguish Use of English tasks from Reading comprehension tasks in the new Reading paper, based on task focus, format and the level of cognitive processes they are predominantly expected to activate. As such, the componential aspect of the exam has been maintained. Despite the reduction in text length and number of items, the paper continues to be robust, allowing for a separate Use of English score to be reported. The close link between linguistic knowledge and reading comprehension supports the merging of the pre-revision Reading and Use of English papers into one paper with two components, as well as the fact that both components are expected to activate cognitive processes along the continuum of the reading comprehension model discussed in Khalifa and Weir (2009). This reading model also supports the positioning of the Use of English tasks at the beginning of the Reading and Use of English paper as they tend to activate the lower-level processes, which eases the candidates into the paper as well as acting as a bridge to the subsequent papers. It is hoped that the revised Use of English tasks will continue to have a positive impact on learning.

Now that the exams have gone live, additional research will be undertaken to monitor both the statistical performance of items and tasks as well as the construct coverage and cognitive processing activated by the tasks.

References

Abraham, R G and Chapelle, C A (1992) The meaning of cloze test scores: An item difficulty perspective, *The Modern Language Journal* 76 (4), 468–479.

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

Ashton, K, Salamoura, A and Diaz, E (2012) The BEDA impact project: A preliminary investigation of a bilingual programme in Spain, *Research Notes* 50, 34–42.

Capel, A (2010) A1–B2 vocabulary: Insights and issues arising from the English Profile Wordlists project, *English Profile Journal* 1, 1–13.

Capel, A (2012) Completing the English Vocabulary Profile: C1 and C2 vocabulary, *English Profile Journal* 3, 1–14.

Celce-Murcia, M and Larsen-Freeman, D (1999) *The Grammar Book: An ESL/EFL Teacher's Course*, Boston: Heinle/Thomson.

Chambers, L, Elliott, M and Jianguo, H (2012) The Hebei Impact Project: A study into the impact of Cambridge English exams in the state sector in Hebei province, China, *Research Notes* 50, 20–23.

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

Docherty, C, Casacuberta, G G, Pazos, G R and Canosa, P (2014) Investigating the impact of assessment in a single-sex educational setting in Spain, *Research Notes* 58, 3–15.

Elliott, M, Docherty, C, and Benjamin, T (2015) *PET(fS) Construct Investigation (SEM)*, Cambridge: Cambridge English internal report.

Elliott, M, Lim, G, Galaczi, E and Calver, L (2012) *FCE and CAE construct validation study (Part 2)*, Cambridge: Cambridge English internal report.

Geranpayeh, A (2007) Using structural equation modelling to facilitate the revision of high stakes testing: The case of CAE, *Research Notes* 30, 8–12.

Geranpayeh, A and Somers A (2006) *Testing the Construct Model for the CAE Examinations*, Cambridge: Cambridge ESOL internal report.

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

Harrison, J (2015) The English Grammar Profile, in Harrison, J and Barker, F (Eds) *English Profile in Practice*, English Profile Studies volume 5, Cambridge: UCLES/Cambridge University Press, 28–48.

Hawkey, R (2009) *Examining FCE and CAE: Key Issues and Recurring Themes in Developing the First Certificate in English and Certificate in Advanced English Exams*, Studies in Language Testing volume 28, Cambridge: UCLES/Cambridge University Press.

Hawkins, J and Filipović, L (2012) *Criterial Features in L2 English*, English Profile Studies volume 1, Cambridge: UCLES/Cambridge University Press.

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.

Malarkey, J and Somers, A (2012) *Testing the Construct Model for the Revised FCE Examinations*, Cambridge: Cambridge English internal report.

Nation, I S P (2001) *Learning Vocabulary in Another Language*, Cambridge: Cambridge University Press.

Nation, I S P (2007) Fundamental issues in modelling and assessing vocabulary knowledge, in Daller, H, Milton, J and Treffers-Daller, J (Eds) *Modelling and Assessing Vocabulary Knowledge*, Cambridge: Cambridge University Press, 35–43.

Purpura, J E (2004) *Assessing Grammar*, Cambridge: Cambridge University Press.

Read, J (2000) *Assessing Vocabulary*, Cambridge: Cambridge University Press.

Shiotsu, T and Weir, C J (2007) The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance, *Language Testing* 24, 99–128.

Somers, A Q (2015) Reporting test scores and the Cambridge English Scale, *Research Notes* 59, 23–31.

Weir, C J (2013) An overview of the influences on English language testing in the United Kingdom 1913–2012, in Weir, C J, Vidaković, I and Galaczi, E D, *Measured Constructs: A History of Cambridge English Language Examinations 1913–2012*, Studies in Language Testing volume 37, Cambridge: UCLES/Cambridge University Press, 1–102.

Weir, C J and Porter, D (1994) The multi-divisibility or unitary nature of reading: The language tester between scylla and charybdis, *Reading in a Foreign Language* 10 (2), 1–19.

Revising FCE and CAE Listening tests

MARK ELLIOTT VALIDATION AND DATA SERVICES, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT
AMANDA CHISHOLM ASSESSMENT AND OPERATIONS, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Introduction

This article summarises the research done to investigate the construct coverage of the *First Certificate in English* (FCE) and the *Certificate in Advanced English* (CAE) Listening papers as part of the revision process of FCE and CAE. The analysis of the FCE Listening component did not indicate any significant issues in terms of construct coverage with respect to the stated aims of the revisions project. Minor changes were introduced, however, which are discussed below.

The analysis of the CAE Listening paper indicated some construct under-representation and the subsequent change – to include a means of testing the ability of candidates to differentiate between points of information at discourse level, i.e. within a speaker's turn – is described below, relating it to Field's (2013) cognitive processing model for listening comprehension.

Changes to FCE Listening

There were four changes made to the FCE Listening paper as a result of research and stakeholder input during the revision process from 2011–13.

Firstly, the range of topics covered within the tests was reviewed in relation to the stated purpose of FCE as being suitable for entry purposes to Further Education. With this in mind, new guidelines issued to item writers stated that 'some tasks will have more of a study focus or work flavour, in line with the exam's use for access to [Further Education] or less demanding [Higher Education] courses, or for work or vocational training purposes. However, it should not be assumed that candidates will have any knowledge of specific workplaces e.g. Human Resources or Marketing or the innermost workings of universities' (UCLES 2013:6).

Secondly, the number of options in Part 3 (the multiple matching task) was increased from six to eight in order to bring it in line with the equivalent task in CAE and reduce the degree of interdependence between items; the task now includes the keys to the five items plus three distractors rather than one.

Thirdly, the extent of the spoken rubrics on the recording was reduced for Part 1, for which previously the context, stem and all three options had been read out (in contrast with CAE, where only the context and stem were read out). This change was introduced for two reasons:

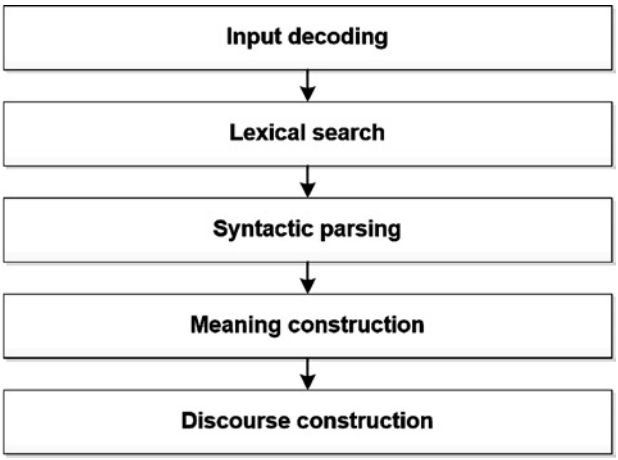
1. To bring Part 1 in line with other parts of the test, where the questions and options are not read out on the recording. There is still a pause in the recording to give candidates time to read the items for Part 1.
2. To shorten the total timing of the test, which was typically longer than that of both CAE and the *Certificate of Proficiency in English* (CPE).

Finally, recordings for Part 2 (the sentence-completion task) were standardised as monologues in order to maximise the usefulness of the text: the brief introductions which some Part 2 recordings contained were untested – that is not required to respond to any of the items – and, since each new speaker requires the listener to normalise to the features of the new speaker's voice (Field 2013:116, Elliott and Wilson 2013:186), a further untested stretch of recording was required, which entails both less efficiency in terms of useful recorded input and a different in test experience for the candidate.

Changes to CAE Listening

In his analysis of the cognitive validity of the CAE Listening component, Field (2013:85–102) investigated the targeting of different levels of cognitive processing within the *Key English Test* (KET), *Preliminary English Test* (PET), FCE, CAE and CPE within his cognitive processing model of Listening comprehension, outlined in Figure 1. In many ways this is analogous to the model for Reading comprehension detailed by Khalifa and Weir (2009) and summarised by Vidaković, Elliott and Sladden (this issue), but naturally contains important differences due to the nature of the input, which means not only that the lower-level processes can take an entirely different form (e.g. the difference between decoding input from an aural stream and a written orthographic representation) but also due to the specific constraints and demands placed upon a listener as opposed to a reader (e.g. the ephemeral, non-standardised nature of the input and the extra working memory demands due to the linear nature of the input).

Figure 1: A model of listening comprehension (adapted from Field 2013:97, 101, 104)



Of particular interest here is the highest level of processing – discourse representation – which describes the process in which the listener 'makes decisions on the relevance of

the new information and how congruent it is with what has gone before; and, if appropriate, integrates it into a representation of the larger listening event' (Field 2013:96). By selecting information the listener determines to be salient (and discarding non-salient information – working memory constraints preclude the integration of all information over an extended time) and incorporating it into a hierarchical structure, the listener creates a high-level *discourse representation* of the text (Brown and Yule 1983:206), into which further information will be incorporated (or discarded). The Common European Framework of Reference (CEFR, Council of Europe 2001) C1 level descriptors for overall listening comprehension state that a listener 'can follow extended speech even when it is not clearly structured and when relationships are only implied and not signalled explicitly' (2001:66) while the C1 level descriptors for listening as a member of a live audience state that a listener 'can follow most lectures, discussions and debates with relative ease' (2001:67); fulfilling both of these descriptors clearly requires the listener to engage in discourse construction. However, Field's investigation into the cognitive demands of CAE Listening found that (2013:138):

> Slightly anomalous, at least in the sample materials, is the fact that Part 4 at FCE features MCQ [multiple-choice questions] items of some conceptual complexity, which oblige the test taker to differentiate quite finely between points of information within an entire speaker's turn (i.e. to operate at discourse level). The same level of informational complexity is not demanded at CAE. At CPE, discourse-level processes are well represented by a task (Section 4) which requires test takers to match opinions against speakers.

This lack of test coverage of discourse construction in some versions of a test primarily targeting C1 level constituted a degree of construct under-representation – a failure to test the full construct to an adequate degree as conceptualised within the theoretical model – which represented a threat to validity. This threat to validity had particular relevance to the CAE Listening component in light of the specified goal of the revisions: to ensure its appropriacy as a test for use as an entrance requirement for Higher Education institutions; Higher Education study involves lectures which require students precisely to assimilate the information presented and incorporate it into a hierarchical discourse representation. For this reason, the consistent inclusion of items focusing on discourse representation became the focal point of the CAE Listening revisions.

Of particular relevance here is Field's (2013:138) observation that 'at CPE, discourse-level processes are well represented by a task (Section 4) which requires test takers to match opinions against speakers'; this observation refers to Part 4 of the pre-2013 CPE Listening component, which involved ascribing stated opinions in a dialogue to one or other of the speakers or to both where they agree. This task involves assimilating each proposition into the discourse representation and determining the relationships (here in terms of speaker agreement).

A similar testing focus to the CPE task, but with less conceptually dense recordings suitable for C1 level candidates as opposed to C2 level candidates, was identified as an appropriate means of covering discourse construction within the CAE Listening component; the key question was then whether the existing task types in CAE Listening could consistently accommodate items with such a focus. A review of the types of recordings in CAE Listening indicated that the only part suitable for accommodating such items would be Part 3, for which the Item Writer Guidelines (UCLES 2007:19) specified two possible types of recording:

> Texts for this part need to be either a) fairly formal interviews with the views of the interviewee forming the basis of the testing points or b) discussions involving two main speakers, both of whose views are tested at different points in the text.

The second recording type (discussions involving two main speakers), which already indicated the testing of individual speakers' views, provided a ready-made platform to more systematically introduce items testing agreement/disagreement across speaker turns, testing discourse construction (although in a different response format from the pre-2013 CPE task: CAE Part 3 which consists of four-option multiple-choice items). The decision was made to discontinue the production of the first type of recording (fairly formal interviews) and specify two-way discussions for future tasks, adding a requirement that items should be included that focus on speaker agreement/disagreement, their shared views or experiences, or testing a speaker's attitude to what had already been said; this relatively minor change to the test specifications achieved the goal. Due to practical considerations, it was decided to phase in the change, with a move to phasing out the older versions of the task type in time. Rather than limiting recording types purely to two-way discussions, interviews with one interviewer and two interviewees were also specified as a possibility, since these make it possible to test speaker agreement/disagreement in an analogous way.

The revised Item Writer Guidelines for CAE Listening (UCLES 2014:28) specify the formats for Part 3 recordings in the revised test:

> Texts will generally feature three speakers, e.g. an interviewer and two interviewees, with long turns from the interviewees providing the tested content. A discussion involving two people (e.g. with different perspectives on the same topic/experience) is a possible variation on the format. Such a discussion might be set up by a presenter who takes no further part in the interaction.

The Item Writer Guidelines go on to specify that items should be included which 'test across turns – for example focusing on areas of (dis)agreement between the main speakers, or on their shared views and experiences' (UCLES 2014:30). Parts 1, 2 and 4 of the CAE Listening component remain unchanged.

Conclusion

The changes made to the Listening components of FCE and CAE were not as significant as those made to some other components; however, the changes which were made in response to both research findings and market feedback serve to strengthen the validity of the tests as being suitable for their stated contexts of use as entrance requirements for Further Education and vocational training (FCE) and Higher Education (CAE) while ensuring continuity with the previous examination format and, in the case of FCE, producing a shorter test in terms of tie without any reduction in test content or items.

References

Brown, G and Yule, G (1983) *Discourse Analysis*, Cambridge: Cambridge University Press.

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

Elliott, M and Wilson, J (2013), Context Validity, in Geranpayeh, A and Taylor, L (Eds) *Examining Listening: Research and Practice in Assessing Second Language Listening*, Studies in Language Testing volume 35, Cambridge: UCLES/Cambridge University Press, 152–241.

Field, J (2013) Cognitive Validity, in Geranpayeh, A and Taylor, L (Eds) *Examining Listening: Research and Practice in Assessing Second Language Listening*, Studies in Language Testing volume 35, Cambridge: UCLES/Cambridge University Press, 77–151.

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.

UCLES (2007) *Item Writer Guidelines: CAE Listening*, Cambridge: Cambridge English Language Assessment internal document.

UCLES (2013) *Item Writer Guidelines: Cambridge English: First: Listening (from 2015 onwards)*, Cambridge: Cambridge English Language Assessment internal document.

UCLES (2014) *Item Writer Guidelines: Cambridge English: Advanced: Listening (from 2015 onwards)*, Cambridge: Cambridge English Language Assessment internal document.

‘Seeing the images with different eyeballs’: Using text-based vs picture-based tasks in the revised CAE Speaking test

NICK GLASSON ASSESSMENT AND OPERATIONS, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

EVELINA D GALACZI RESEARCH AND THOUGHT LEADERSHIP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Introduction

In late 2011, Cambridge English embarked on a revision of the *Certificate in Advanced English* (CAE) exam (targeted at CEFR C1 level), in line with the regular exam review cycle which Cambridge English exams go through. A global aim of the review was to consider ways in which the Speaking component could be revised and improved. Specific aims were to focus on task features which are part of the ‘context validity’ of the task (Weir 2005), such as the timing allocated to the task, the nature of the prompts used, and the content of prompts. Such context validity considerations play a key role in determining the type of language generated by learners during the test, and are therefore a key consideration in the development of a new or revised test.

At the time of the revision the exam included four parts and the following task features:

- Part 1: a question-and-answer task with *spoken prompts* (questions) delivered by the examiner (3 minutes)
- Part 2: an individual ‘long turn’ task with a combination of two written question prompts and a selection of three visual *picture prompts* and one further spoken prompt (question) delivered at the end of each turn to the Listening test taker (1 minute per test taker, 4 minutes overall)
- Part 3: a paired discussion task with a set of visual *picture prompts* organised around a common theme and with two written question prompts (4 minutes)
- Part 4: a three-way discussion task involving the two test takers and examiner with *spoken prompts* (questions) delivered by the examiner (4 minutes).

The focus on the nature of the prompts was driven by the importance of ‘channel of communication’ as a contextual task validity feature (Galaczi and ffrench 2011, Weir 2005). Channel of communication refers to the way the task is communicated, which in a Speaking test is typically aurally through questions spoken by the examiner or visually as text-based or image-based prompts. Research has indicated that the nature of the prompt is not insignificant and could play a potentially important role in determining the type of language elicited from the test takers. In this respect, O’Keefe (2006) found that test takers responding to text-based prompts produced more complex language than when responding to visual image-based prompts, and noted that visual prompts could constrain language since test takers would discuss the visuals themselves rather than the concepts they represented, often with pre-prepared phrases (e.g. ‘Here we can see . . .’), which limited the complexity of their language.

Following a review and consultation exercise involving the Chair of the speaking component and other key personnel connected to CAE Speaking, certain test design features were incorporated into draft test specifications and trialled. The CAE trialling decisions were situated within a wider project of revising the *First Certificate in English* (FCE), *First for Schools* and CAE, which were undergoing revision at the same time, and collaborative discussions about trialling decisions across these three exams were a prominent feature of the revision project.

Draft test specifications: Task design variants

For Part 1 of the test, a reduced set of introductory questions was decided on for trialling. The main impetus for this decision was drawn from the idea that while the overall test length remained unchanged, a shorter Part 1 could allow time

to be allocated in the more challenging parts of the test, thus allowing advanced-level candidates a better opportunity to show what they could do linguistically. This decision was supported by the expert judgement of the revision team, who felt that a shorter Part 1 would still allow for coverage of the testing focuses of that task (i.e. general interactional skills) and therefore would not compromise construct coverage. The Part 1 questions were selected as general interactional prompts, such as 'What has been your most interesting travel experience?'. A further rationale supporting the decision for a slightly shorter Part 1 was the decision to make the CAE exam more similar in design to the *Certificate of Proficiency in English* (CPE) Speaking exam, thus leading to a stronger family resemblance between these two C level exams.

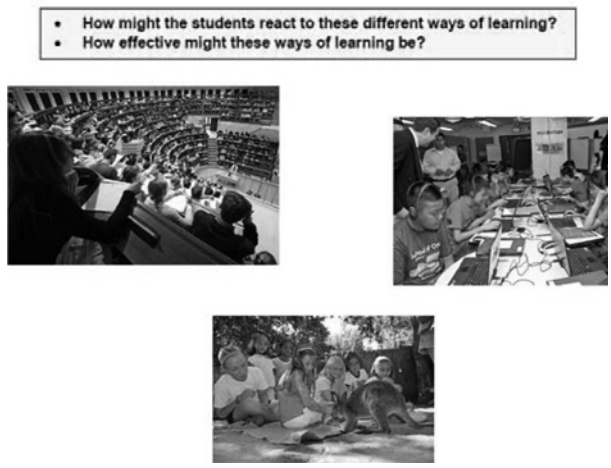
In the Part 2 'long turn' task, a text prompt was developed (Figure 1) alongside the existing picture prompt (Figure 2). This alternative prompt format was developed as a direct response to the findings of O'Keefe's (2006) study, which indicated that text prompts might provide a richer sample of language at the C1 level. The text-based task design was also similar in scope and demand to that of the long turn in CPE, thus leading to a family resemblance between task types in the C level exams of the suite.

Figure 1: Text-based Part 2 task

'Everybody has the ability to do any job.'
What do you think?

- training
- motivation
- personality

Figure 2: Existing visuals-based Part 2 task

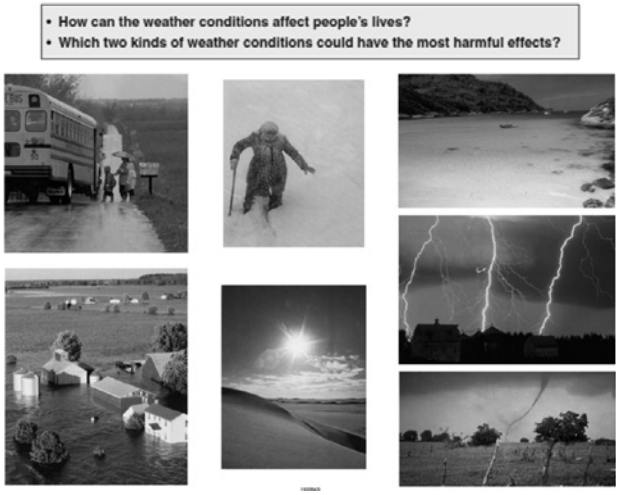


For both the alternative (text-based) task prompt (Figure 1) and the existing (picture-based) task prompt (Figure 2), a timing of 1 minute and 30 seconds was allotted for the 'long turn' response, which was longer than the existing 1-minute timing for this part. The rationale for this decision was not based on any empirical research, since there is no research specifically investigating optimal timing of different task types at different CEFR levels. Rather, it was based on the collective expert judgement of the revision group who felt that a long turn longer than 1 minute at the C1 level would provide test takers with more opportunity to display advanced language skills. In addition to the longer time for this task, a 30-second preparation time was allocated, in line with suggestions provided by Field (2011) about planning time for tasks which involve long turns.

As noted in the overview of the test format above, this task also contained a question for the listening candidate. In the revised test specifications, this question was altered from a task-specific question (e.g. 'Candidate B, which of these situations would be the most memorable? . . . Why?') to a generic one, regardless of topic or theme (e.g. 'Candidate B, what do you think?'). This decision was driven by the dual considerations that a more open-ended question would be more suitable at a C level, and the higher practicality of producing tasks with generic prompt-neutral questions, which can be used across tasks.

Part 3 of CAE was the area where there was the greatest potential for change. The existing task consisted of two questions and a range of thematically related visuals intended to convey a particular 'angle' or idea in relation to the questions, as seen in the example in Figure 3.

Figure 3: Existing visuals-based Part 3 task



The pilot version of the task was initially designed with text prompts listed in the manner shown in Figure 4. After initial trials an alternate design was also used, a diagram, as shown in Figure 5. The main reason for the change in design was the authenticity of a mind map, which learners (and by extension CAE test takers) would often use in their educational settings.

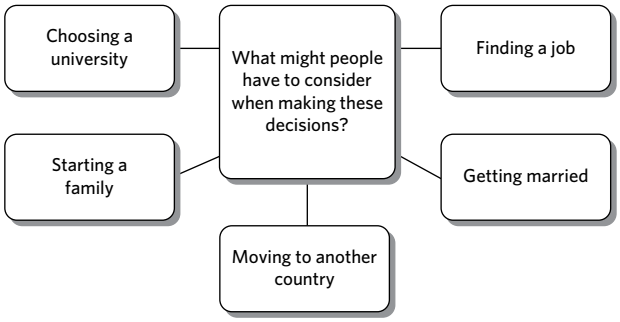
Figure 4: The 'list' task design used initially for the text-based Part 3

What are the advantages and disadvantages of finding out about the past in the following ways?

- Visiting museums
- Using the Internet
- Archaeology
- Historical re-enactments or role plays
- Visiting historical sites and buildings

A further change made to this task was a reduction in the number of ideas provided, from six or seven options (as seen in Figure 3) to five options (as seen in Figures 4 and 5). This change was motivated by feedback from examiners

Figure 5: The 'diagram' task subsequently designed for the text-based Part 3



and Chairs that a large number of ideas provided in the task could lead to superficial interactions between the test takers with limited topic development, triggered by their attempt to cover all ideas provided. Limited topic development could, in turn, fail to provide test takers with enough opportunities to fully display advanced interactional skills, which involve co-constructed topic development across several turns, as shown in Galaczi's (2014) analysis of learner interactional skills. This need to provide test takers freedom to develop topics at length across multiple interactional turns had to be balanced with the need to provide content scaffolding for test takers through the inclusion of a range of ideas, not all of which have to be necessarily discussed in the task. As a result, five options were included in the task.

An additional feature to the text-based prompts in Part 3 was, as with Part 2, the inclusion of preparation time to allow candidates a short window in which to process input before speaking. This decision was partially informed by cognitive validity considerations and the need for some planning time in long turn responses, as discussed by Field (2011). Even though Part 3 does not entail long turns, it was felt useful at the trialling stage to explore the inclusion of planning time in this part.

A final task feature to be trialled was the introduction of a 'split rubric' (i.e. task instructions) for the Part 3 collaborative task, whereby the *discussion* and *decision* focus of the existing task would be separated, as shown in the draft interlocutor frame in Figure 6.

In line with the decision to use text prompts, the rationale for using a split rubric was initially based on a finding identified by O'Keefe (2006:18) as a way to make this paired task more focused. As O'Keefe suggests, the inclusion of several task focuses (e.g. discuss and then decide) for a 3-minute collaborative spoken activity requires that 'one of the candidates assumes a higher level of control over interaction than their partner' in order to manage the interaction and ensure that all task requirements are met. This finding was also evident in filmed samples of CAE and was an issue that many Speaking Examiners had commented on during operational administration of the Speaking test, namely that candidates are often reluctant to assume this 'power holding' role in a collaborative task and that 'topics meander', leading to poor task completion. A staged approach to this task, where the task requirements are more distinctly split and the time is correspondingly divided was felt to make the interaction more manageable and focused.

Part 4 of CAE was not considered to need significant revision as it was generally seen as effective. The main change related to the topics included in it and the general intention for it to be further removed from the topic in Part 3 in order to widen the discussion and reduce the risk of overlap between Parts 3 and 4 in candidate responses. The timing of this section was extended in the light of shortening the Part 1 phase of the test and an additional question was added to reflect this longer timing. This was felt to allow candidates to provide a more suitable sample of language at this proficiency level.

With these task variants established trialling began in February 2012.

Trialling

Methodology

The trialling methodology involved two phases, in which both text- and picture-based task variants were trialled. The first phase focused on the task variants individually, with a view to gathering feedback on how the two different prompt versions were performing and without considering how they would fit into the full test. The second phase involved trialling different versions of full tests.

In both phases, there was a consistent drive to put the test takers at the centre of discussions and a considerable amount

Figure 6: Interlocutor frames for split rubric (early draft version)

Part 3

Interlocutor Now, I'd like you to talk about something together for about two minutes.
(3 minutes for groups of three)
Here are some descriptions of ways in which people can find out about the past.
*Place **Task 21**, in front of the candidates.*
Talk to each other about **the advantages and disadvantages of finding out about the past in these ways**.
You now have up to 15 seconds to prepare.
15 seconds
All right? Would you start now, please?
2 minutes (3 minutes for groups of three)

Candidates

Interlocutor Thank you. **Now you have a minute to decide which of these ways of finding out about the past would be most popular with young people**.
Approx. 1 minute (2 minutes for groups of three)
Thank you. (Can I have the booklet, please?) *Retrieve **Task 21**.*

of time was spent getting as much feedback from the trial test takers as possible. As such, the trial consisted of several complementary strands: test taker feedback from (often quite extensive) discussions following the trials and observer notes; analysis of language functions in the test taker language generated by the different prompts; a comparison of test taker scores generated by the different prompt versions.

A total of 24 learners participated in Phase 1 of the trials and 28 in Phase 2. The trial was part of a larger project focusing on changes in the task prompts in the FCE and FCE for Schools exam, which included a similar number of participating test takers.

The participating learners were selected to offer a range of first languages, ability levels, and test preparation experiences, and thus to be representative of the test taker population. The examiners involved in the trials provided a rough estimate of the linguistic ability level of the participating learners in relation to CAE, and in Phase 1 12 (50%) were judged to be 'average', 10 (42%) either 'average to strong' or 'strong', and two (8%) 'weak' and slightly below C1 level.

The candidates were mostly preparing for the CAE exam, a minority were preparing for FCE (CEFR Level B2) or CPE (CEFR Level C2), and some were taking both CAE and CPE. The majority of test takers fell within the age range of 18–25 years. The smallest amount of time spent studying English was three years, with the majority reporting that they had been preparing for the exam for 1–2 months. This learner profile was consistent across both trial phases.

The trials were conducted by the Assessment Manager, Chair, an item writer and two experienced Speaking Examiners (who alternated between the roles of Interlocutor and Assessor). As such, the team involved in the trials brought in expertise from different perspectives. Observation and feedback forms designed specifically for the trials were completed by all observers and raters, and, with participant permission, interactions were recorded (with the exception of one interaction). All students participating in the trialling session signed media recording release forms.

The analysis involved a thematic analysis of test taker and examiner feedback, a functional analysis of linguistic functions in the language generated by the revised prompts, and a statistical comparison of scores.

Trialling: Phase 1

During the first trialling phase, every test taker was given a full test (based on the current test format), and additionally the text-based prompt versions of Part 2 and Part 3, with a view to gathering feedback and observations on those task variants in isolation, rather than looking to how the new task versions fit into the wider test.

The context of the trialling was explained to each pair of test takers. Then one of the Part 1 sets was trialled. This was followed immediately by both a text-based prompt and a picture-based prompt for Part 2. The trial then paused to capture feedback, and resumed with both Part 3 tasks (one text based and the other picture based) and again, after the Part 4 segment had concluded, students were asked for their views. The order of picture and text versions was alternated each time, to reduce an order effect.

In Phase 1 of the trialling there was an emphasis on determining what basic linguistic functions the task designs were eliciting in use. Consequently, observers were asked to quantify instances of functions such as 'speculating', 'justifying opinion', etc. from a set list of options while tasks were trialled. The data from this is presented below, where relevant to the discussion.

In Phase 1 preparation timings for text-based Part 2 and Part 3 tasks were experimented with (within a range of 15–30 seconds), which informed the timings set at task level for Phase 2 (Table 1).

Table 1: Outline of test designs in Phase 2 of the revision project

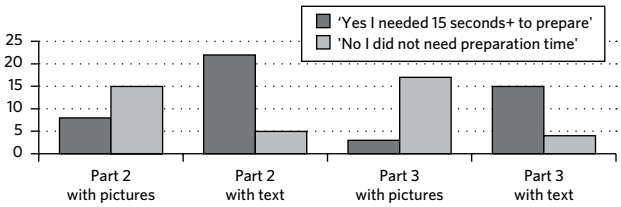
Design	Part 1	Part 2	Part 3/4	Total time
A	6 questions 2 minutes	Text prompt tasks 30 second preparation time 1 minute 30 second 'long turn' Question for listening candidate + a 30 second listening candidate response 5 minutes	Picture prompt tasks No preparation time Split rubric structure 2 minutes: discussion 1 minute: decision Follow-up questions 8 minutes	 15 minutes
B	6 questions 2 minutes	Picture prompt tasks No preparation time 1 minute 30 second 'long turn' Question for listening candidate + a 30 second listening candidate response 4 minutes	Text prompt tasks 15 second preparation time Split rubric structure 2 minutes: discussion 1 minute: decision Follow-up questions 9 minutes	 15 minutes

At the end of Phase 1 of the trials a complete review was carried out of the tasks, their design and their relative merits, based on test taker feedback, observer notes and analysis of functions generated, in order to inform Phase 2 of the trialling.

Trialling: Phase 2

Based on the insights gained in Phase 1 it was recommended that trialling for Phase 2 would be carried out on full tests and that visual prompts should be retained *somewhere* in the test, in order to allow variation in prompts, as recommended in Galaczi and ffrench (2011). Phase 2 would, as such, incorporate trials of two whole tests (Table 1) so that further analysis of Part 2 and 3 text-based tasks could be done alongside the rest of the revised test content. The allocation of preparation timings in Phase 2 was the direct result of feedback from learners (as summarised in Figure 7), who felt there was a need for preparation time for text-prompted tasks rather than visuals-based tasks.

Figure 7: Learner feedback on necessity of planning time (N of test takers)



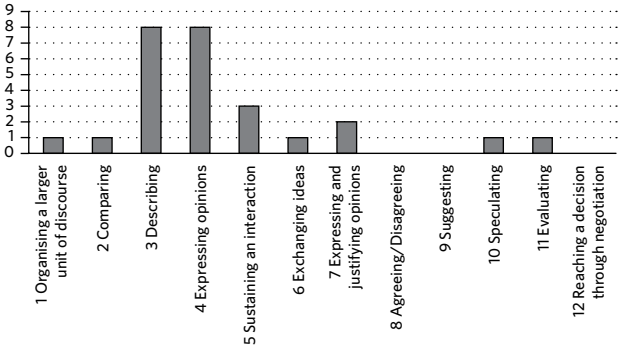
Findings

Our discussion of findings will first briefly focus on the two parts which underwent the least change – Part 1 and 4, and will then move to a more extensive discussion of Part 2 and 3.

Part 1 (Question and answer): Reduced number of questions

In Part 1 the test takers were given one or more questions depending on the response they gave to the initial general questions in this part and the time taken to answer, which provided a preliminary initial gauge of their ability.

Figure 8: Linguistic functions observed for Part 1 (with a reduced number of questions) in Phase 1 (N=24 test takers)



As seen in Figure 8, the main functions observed during Phase 1 trialling for Part 1 tasks were 'describing' and 'expressing opinions'. This is consistent with the intended focus of this part of the test, which aims to tap into informational and interactional functions such as describing, expressing opinions, justifying opinions, and exchanging ideas. Feedback from test takers indicated that the questions in Part 1 gave them 'enough to talk about' and were 'interesting'; comments from observers were positive overall, with statements such as 'gave a full answer' and 'seemed interested in the topic' being typical comments.

Part 4 (Three-way discussion): Including an additional question and specified timing

As has already been noted, there were no significant plans to change the final section of CAE Speaking. However, based on the trials, there was a move to including six, as opposed to five, questions in this part. Similarly, there was a decision to clarify the timings for Part 3 and Part 4 where previously there had just been an overarching time given. It was felt that the Part 4 would benefit from an additional question in terms of allowing for greater elaboration, extension and exploration of a theme while also allowing examiners more options when directing questions to one or more test takers. This was something noted in trialling as a positive move, one that generally helped CAE bridge the concrete-abstract span between FCE and CPE.

Part 2 (Long turn): Text vs picture prompts

The text version Part 2 tasks were trialled alongside the picture-based format, which includes a set of three pictures and two overarching questions (as shown in Figure 2).

One task feature under investigation in this task was whether the ideas included as bullet points provided a sufficient amount of input for test takers to talk for the time required. The feedback from the trial indicated that the bullet points and questions provided were sufficient, although some weaker candidates felt they were not given 'enough to talk about'. Most candidates used two or three of the bullet points provided to help them frame their long turn.

A further feature of interest was inclusion of planning time in the text-based task. Test takers commented that they felt the 30-second preparation time given was sufficient.

Regarding the length of time for completing the task, in both phases of the trials, and with both prompt formats, learners were found to struggle to fill one and a half minutes for an individual long turn. Two main reasons were found to be causing this: some learners were so used to the 60-second task length of the existing format that out of habit (and training) they stopped after a minute, while others found it difficult to keep finding things to say. It is difficult to quantify how many learners failed to fully sustain a 1 minute 30 second long turn, as some learners came to an abrupt end while others were more able to talk at length but relied on pauses and repetition to make it through. The general consensus of those present in the trials was that adding 30 seconds to the long turn did little to improve the quality of the language sample generated or the test experience for candidates.

Observers noted that the text prompts used in Part 2 were more prone to pauses and hesitation. This was echoed in learner feedback, which noted that the 'pictures give you more to say' and allow you 'more choice in how to deal with the task'. Some comments related to the text prompts also pointed to the rather more abrupt nature of the bullet-point design generally, with one learner remarking that with the text prompts used for the long turn they 'didn't know where to begin'.

In light of comments from learners, direct observations and

trial recordings, the group of experts involved with the revision and trialling decided that if a test taker struggles to fill their time on the individual turn in Part 2, then it was preferable to use the visual prompts, since the images were a more 'open' resource of ideas than the text prompt design used in these trials. The general opinion was that weak candidates especially would at least be able to describe the image content whereas with text the weaker candidates would be more likely to stall or become nervous, which could negatively impact on Part 3 too. While CAE aims to go beyond the level of picture description in order to provide opportunities for test takers to display higher-level linguistic functions, it must be recognised that weaker candidates may need to rely on this more basic language function to sustain their long turn.

Figure 9: Linguistic functions observed for Part 2 with text prompts in Phase 1 of trialling (N=24 test takers)

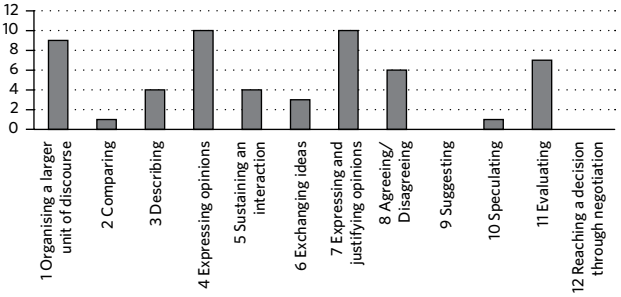
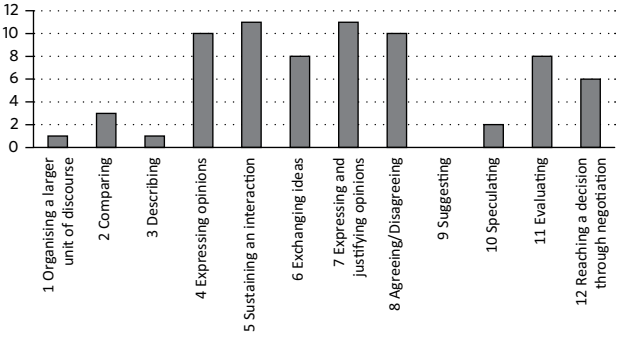


Figure 9 provides an overview of the language functions generated in Part 2 in Phase 1 of the trials, and indicates that a range of linguistic functions were generated. The main functions observed with the text prompts were 'expressing opinions', 'expressing and justifying opinions' and 'organising a larger unit of discourse'. This is in line with the expected test focuses in this part of the test, which is framed with a 'What do you think?' question.

Part 3 (Paired discussion): Text vs. picture prompts

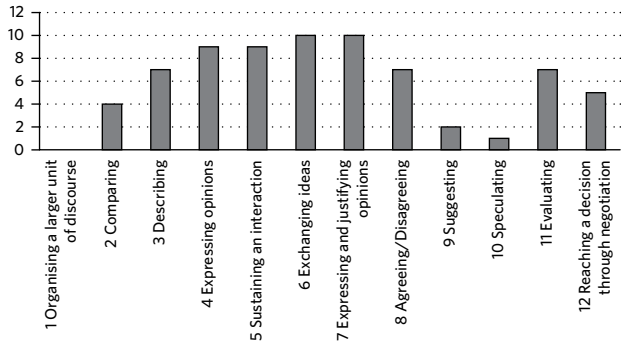
As noted earlier, Part 3 involves a discussion between test takers, with the aim to elicit interactional language functions. While the distribution of language functions for Part 3 was broadly the same across the two prompt types, a difference was observed with 'describing', which was lower in the text prompt observational data (Figure 10) than in the picture prompt (Figure 11).

Figure 10: Linguistic functions observed for Part 3 with text prompts in Phase 1 of trialling (N=24)



The following Part 3 discussion openings illustrate this (references to visuals are in bold):

Figure 11: Linguistic functions observed for Part 3 with picture prompts in Phase 1 of trialling (N=24)



Excerpt from text-based prompt 'Finding out about the past'

Interlocutor:	First, talk to each other about the advantages and disadvantages of finding out about the past in these ways. All right?
Candidate A:	If I start us off . . . I would say the most important advantage is in the historical re-enactments or role plays and visiting historical sites and buildings because your mind really gets involved . . . it's more interesting, more intriguing than visiting museums which is quite well let's say dry for me . . .
Candidate B:	Why do you say that?
Candidate A:	Because you can just go through the rooms and watch certain things that are in this exhibition and you can think about it, you can read about it in descriptions but . . . your mind can't really get involved like in the re-enactments where you have to follow a story.
Candidate B:	Well I think that it depends . . . for example, if you're going to a museum to do an exhibition, to do a painter exhibition you can do it only in a museum, you can't do it in an historical site.

Excerpt from picture-based prompt 'Finding out about the past'

Interlocutor	First, talk to each other about the advantages and disadvantages of finding out about the past in these ways. All right?
Candidate A:	Okay.
Candidate B:	Okay.
Candidate A:	Err . . . **the first part is a museum I think**
Candidate B:	**Yes, it's a coliseum? Is that coliseum?**
Candidate A:	**Yes, coliseum, natural history museum . . .**
Candidate B:	**Yeah**
Candidate A:	Well, the advantage is of course that everything is pretty close by . . . in your own country – so you don't have to travel far and well you get a lot of information I think, you?
Candidate B:	Yeah and finding about the past **this way** and erm . . . finding [unclear word] gives you the opportunity to know about the first lives in the world . . . and the disadvantage could be that I think it's rather complicated to put all the pieces together.

Candidate A: Yeah, museums got kind of a dull image. I think it's more perceived to be for older people and stuff like that . . . and when you for example, compare it to going to . . . **I think it's the acropolis right?** . . . going to a history site I think it's more . . . appealing to people I think.

Some of the comments made by observers in the trials also highlight the difference between the prompt types: observers noted that 'better language was elicited on this text-based task' and that the 'discussion(s) seemed to go better' with 'no speculation involved in trying to decide what pictures showed'. It was observed that in the text-based Part 3 candidates 'got straight into discussion'.

A further focus of the trial was the use of a split rubric and associated timings in Part 3. The split rubric worked well throughout the two phases of trialling in terms of giving the candidates a much clearer structure to follow, negating the need for them to take on an interlocutor-like role and manage the discussion. The participating test takers and observers unanimously commented that they preferred the split-rubric structure, as it represented two distinct *discussion* and *decision* stages, which made the task clearer and more focused. Both observers and test takers commented that splitting the task into two elements provided a more clearly guided and structured activity, which at the same time did not lower the difficulty of the task. More clarity and focus in task rubrics makes it more likely that all intended test focuses would be covered, thereby enhancing the validity of the task. It was also noted that the use of the split rubric can help to redress the balance when one speaker tends to dominate the discussion, as it creates a break and a chance for the more passive speaker to re-establish some participation in the interaction.

A few concerns were expressed about the split rubric. A key one was that some candidates might make a decision in the first phase of the task, which would render the decision phase obsolete. The revision team members felt that this was a matter to be addressed in the way tasks are written through moving the decision question away from the discussion so as to avoid this potential overlap.

The timings associated with the split rubric were also investigated throughout trialling. The allotted 2 minutes for discussion followed by 1 minute to reach a decision seemed appropriate and this was consistently shown to be suitable for the CAE candidature. Both the revision trial phases and subsequent trials for live materials have shown that the timing is adequate.

In addition, a preparation time of 15 seconds for the text prompts was felt to be suitable for the Part 3 task. Both students and observers noted that some time to process the text content was needed, which supported the decision to include a short preparation time in the task. At the same time, test takers commented along the lines of 'I need 15 seconds but 30 seconds is too long'. The length of preparation time – 15 seconds – was an attempt to balance the test takers' need to process the text content and the test validity need to build interactional authenticity into the task through lack of planning time, which is typical for extemporaneous speech (Field 2011). Providing excessive planning time in a discussion task could potentially stifle genuine spontaneous interaction and lead to test takers taking turns to give short 'speeches' rather than engaging in a co-constructed discussion. This decision was subsequently corroborated by research by Nitta and Nakatsuhara (2014), who raised concerns that planning time in interactive tasks might limit test takers' opportunity to demonstrate their abilities to interact collaboratively.

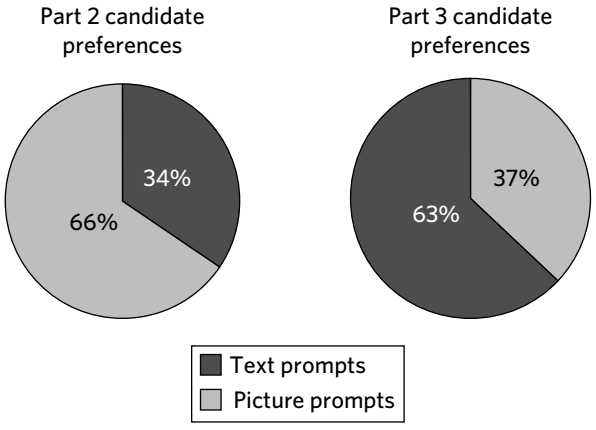
A final point of interest regarding Part 3 was the list-based task design (Figure 4) versus the mind map diagram design (Figure 5). It was found that the listing of options tended to dictate the order of the discussion to a much greater extent than it did with the mind map design. In order to avoid this imposing of a prescribed order for the interaction, and to better mirror the picture-prompted task design it was replacing, the mind map design was adopted. The overwhelming message from trial participants was that they preferred the mind map to the list format. Not one candidate consulted said that they preferred the list design and all either said they would prefer the mind map design or had no strong opinion either way. Comments relating to the mind map raised the following points:

> The mind map looks clearer, easier to read, use and share.
>
> The way the mind map is designed allows candidates to determine their own order.
>
> The mind map is more visually engaging.
>
> The list format does not refer/focus you back to the discussion question as well.

Part 2 and Part 3: Differences in Text-based prompts

A general finding to emerge related to the difference in performance of the text-based prompts in Part 2 and Part 3. In a text-based Part 2, more hesitation and repetition was generally observed. This was, in contrast, not observed with the text-based Part 3 prompts. This view was echoed in test taker feedback in response to the question 'Which format did you prefer? Words or photographs?'. Figure 12 presents the summary of preferred prompt type by test part.

Figure 12: Test taker preference: Text vs. picture prompts



Participants gave various reasons for preferring text in Part 3, but the main theme across their feedback related to the importance of having clear task input, which they felt text can do better as it is more specific and leaves less room for interpretation. In a discussion task using image-based prompts there is a potential risk of misunderstanding of the picture prompts which could affect the interaction. As one student put it: '. . . we're seeing the images with different eyeballs'. This validity threat is minimised in the Part 3 image-

based prompts through careful editing of the task content and trialling. A text-based variant minimises the risk of misinterpretation between speakers even further.

This comment aptly points to the main impetus for retaining visuals in the individual long turn (Part 2) task but replacing them with text in the discussion task (Part 3). The images used in the Part 2 long turn task are intended to give a general context for responding to the two questions about the images, whereas the same is not true of Part 3, where the images used are not intended to provide a defined context *per se* but are much more focused on conveying specific ideas about the topic. For example, in a picture-prompted Part 3 which required the candidates to talk about different people's roles in society and their relative importance, one of the images might be a doctor or nurse. In such cases it is usually intended that the discussion be focused on the general importance of doctors/nurses to society and not the specific details of the context or person in the photograph.

Score distribution comparison between the different prompt types

Scores were awarded for each of the pairs which were successfully recorded in Phase 1. In every case the recordings were rated twice by experienced CAE examiners. The score analysis indicated that there was not a significant difference in the scores achieved in either prompt version (Table 2), with the exception of Interactive Communication. The score difference for Interactive Communication is possibly explained by the fact that some task prompts led to lower interactivity, which would have been reflected in the score for Interactive Communication.

The similar mean scores would suggest that, statistically at least, there was little difference in the scores and language produced by test takers whether they were given text or picture prompts. The issue of text versus pictures, therefore, was not so much about performance, which was similar in terms of language functions and scores, but how best to support the test takers in addressing the tasks and how best to tailor the tasks to balance the concerns of candidates and examiners alike.

Trial outcomes

Part 1

Part 1 was felt to operate well with a reduced set of questions on general topics. In the final test specifications, a set of eight questions for examiners to choose from was adopted for 2015. This aimed to provide examiners with enough material to sustain the short interactions which Part 1 aims to elicit.

The overall timing for Part 1 was also reduced (to 2 minutes for pairs and 3 minutes for groups of three), given the smaller provision of questions in Part 1. This allowed for the Part 4 timing to be increased, as it was felt that a longer Part 4 would be a more valuable means of providing a suitable language sample, since the demands of the questions in Part 4 were more useful as a measure of ability at the C1 level.

Part 2

The idea of completely removing images from CAE Speaking was considered, based on the score comparability between the text and prompt versions, and the higher practicality of producing text-only tasks. However, it was felt that this would considerably diminish the validity of the exam and its appeal to both learners and teachers. From a construct validity perspective, such a change would have affected the contextual validity parameters of the test, as discussed by Galaczi and ffrench (2011), who noted that the *range* of channels of communication in the Cambridge English Speaking tests (i.e. the medium through which the prompt is presented) contributes to the construct validity of the test. A change to text-based prompts in Part 2 of the test would have resulted in a reduced number of types of prompts, possibly affecting the context validity of the test. Such a change may also have had a negative washback effect, and led to a whole raft of common and useful classroom-based test preparation activities involving images becoming redundant and potentially affecting learner engagement in class. As a result, Part 2 has retained the two questions plus three picture format for 2015.

An additional consideration supporting this decision was concerns about the ability of weaker candidates (at both CAE and FCE levels) to deal with the demands of the text input. There were, additionally, concerns about the degree of overlap between a text-based Part 2 and the Part 4 questions. From a test construction perspective, this would be problematic given that the topics and themes which lent themselves to the text-based Part 2 in trials were very similar in area and scope to those of Part 4. This was, indeed, a theme noted in the feedback from both candidates and examiners, who felt that there were similar ideas in play. A final consideration informing the decision about retaining the picture-based task design in Part 2 was a fundamental concern to ensure the new test

Table 2: Score comparison between different prompt versions (15 pairs, 30 test takers)

Assessment category	Prompt version	Mean	Standard Deviation	Z	Sig. (2-tailed)
Grammar	P2 picture/P3 text	2.9	0.4	–1.2689	0.20408
	P2 text/P3 picture	3.1	0.4		
Vocabulary	P2 picture/P3 text	3.0	0.5	–0.0188	0.98404
	P2 text/P3 picture	3.2	0.4		
Discourse management	P2 picture/P3 text	3.0	0.5	–1.4421	0.14986
	P2 text/P3 picture	3.2	0.5		
Pronunciation	P2 picture/P3 text	3.2	0.5	–1.695	0.09102
	P2 text/P3 picture	3.2	0.8		
Interactive Communication	P2 picture/P3 text	3.5	0.8	–3.0544	0.00228
	P2 text/P3 picture	3.3	0.5		
Overall	P2 picture/P3 text	3.2	0.5	–1.1212	0.26272

specification was comparable with its predecessor. Adopting a text-based Part 2 would arguably have involved a more fundamental change of the test specifications and construct, away from the stated focus of 'expressing opinions through comparing' which the existing test specification stipulated.

In terms of timing, the trial findings indicated that a 1 minute 30 second 'long turn' in CAE was often problematic for test takers. As a result, the CAE 2015 Part 2 has retained the timing of 1 minute. While this allows a strong candidate to provide a suitable sample of language, at the same time it is not too long as to make weaker students feel deficient in some way.

Part 3

Part 3 underwent some significant changes following the trials. The new test design has adopted text prompts in a mind map design (Figure 5) and a split rubric in line with feedback and observations from the trials. The removal of images aimed to support a more focused interaction, since the text provided a more specific idea than an image. A mind map comprising five text prompts was decided on, based on the data gathered, suggesting longer developments of topics.

Part 4

More time is apportioned to this part of the test as a result of the revision, as well as a stipulated timing for this test part. Previously the timing for Part 3 and 4 was given as a whole, but in an attempt to further standardise examiner delivery a by-part timing was introduced. Another minor alteration has been the instruction to examiners to use the questions 'in order' but 'as appropriate'. This was reworded slightly to reinforce the notion that the questions in Part 4 are designed to expand the scope of the discussion as they are used. While it is still possible for examiners to leave out questions they feel have already been covered, this was another attempt to help standardise the test process for both examiners and candidates, while at the same time allowing some flexibility.

Conclusion

Following the revision of the test specifications based on the trial findings, tasks have gone through regular rigorous operational trialling procedures. From 2012 to the time of writing this article, over a hundred Part 3 tasks using text prompts and with mind map designs have been created to the new test specifications, of which 52 have gone through operational trialling and, following standard task review procedures, have been selected as suitable for live use. Overall, the adoption of the revised format has been a positive experience for the learners involved, as seen in the trial candidate feedback.

The CAE Speaking revision project trials and the resulting decisions taken about the test specifications were informed by a range of sources, which brought in different perspectives about the test: the expert judgement of assessment specialists and the views of learners; statistical information on test scores and qualitative information on learner language; observer notes and interviews with learners. Throughout this process, the test taker was always at the centre, and indeed, some decisions, such as the use of a mind map in Part 3, were made with direct input from test takers. Such an approach addresses concerns that test designers focus merely on the test 'instrument', where 'those who take the test . . . seem less important' (Underhill 1987:3). Extensive space was given during this project to allow trial candidates to inform the development of tasks and this is hopefully reflected in the final test specifications for 2015, which can be found on the Cambridge English website (www.cambridgeenglish.org/exams/advanced).

References

Field, J (2011) Cognitive validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 65–111.

Galaczi, E D (2014) Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics* 31 (3), 307–328.

Galaczi, E D and ffrench, A (2011) Context validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 112–170.

Nitta, R and Nakatsuhara, F (2014) A multifaceted approach to investigating pre-task planning effects on paired oral test performance, *Language Testing* 31 (2), 147–175.

O'Keefe, A (2006) *Certificate in Advanced English: Evaluation of Modifications to Speaking Test*, Cambridge: Cambridge ESOL internal report.

Underhill, N (1987) *Testing Spoken Language: A Handbook of Oral Testing Techniques*, Cambridge: Cambridge University Press.

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

Aspects of the revision process for tests of Writing

GAD S LIM RESEARCH AND THOUGHT LEADERSHIP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Introduction

The revised *First Certificate in English* (FCE), *First Certificate in English for Schools* and *Certificate in Advanced English* (CAE) exams introduced in 2015 saw the inclusion of compulsory essay tasks in Part 1 of the Writing papers. The genre was one option among others in the previous versions of these exams. Because a growing number of candidates use these exams for further and Higher Education purposes (Howden and Mehta, this issue), there was a need to ensure that ability to write in this academic genre was tested. As before, candidates are able to select from a modified range of genres in Part 2 of the papers.

The essay tasks were developed to elicit evidence of writing ability at Common European Framework of Reference (CEFR) Levels B2 for FCE and FCE for Schools and C1 for CAE, respectively. At the C1 level for example, a learner should be able to 'write clear, well-structured expositions of complex subjects, underlining the relevant salient issues' and to 'support points of view . . . with subsidiary points, reasons and relevant examples' (Council of Europe 2001:62). The CAE essay task (see the Appendix for a sample) thus asks candidates to address relatively abstract/conceptual topics, so as to elicit writing about *complex subjects*. Asking candidates to select the more important of two or three considerations allows them to take a *point of view*/to demonstrate ability to *underline relevant salient issues*. Discussing the options should result in candidates providing *subsidiary points, reasons or relevant examples*, but in order to make it explicit the task asks them to give reasons in support of their answer.

Development and validation of the essay tasks followed the socio-cognitive approach to test validation (Weir 2005), which sets out the different aspects for which validity evidence needs to be obtained. The validation process involved repeated rounds of trialling and multiple sources and types of data. In this article, I highlight some aspects of the development and validation process. For ease of exposition, I primarily exemplify each of the points using just one test rather than using all of them.

What should the task be like?

Beyond providing some input for candidates to respond to, many other features of the task needed to be considered. It even extended in this case to determining which words in the instructions to render in bold, so as to provide candidates further cues on what the task expectations are! Of more consequence, a major focus of the trialling was on the number of words candidates should produce and the amount of time they should be given to do this.

In the first round of trialling, the parameters for FCE and FCE for Schools were set at 120–170 words per task, with 60 minutes given to complete the two tasks. The data indicated that these parameters were probably inappropriate. On average, candidates produced 176 words for the essay task, and in their feedback, about a third of them indicated that the word count guideline did not allow them to fully demonstrate their B2 level abilities.

The time given for the paper was raised to 70 minutes and subsequently to 80 minutes, and the word count guideline to 140–190 words. Consequently, trial candidate perception of the test improved (Table 1). Open-ended feedback related to not having enough time or not having enough words also decreased by a large amount, though the latter remained the subject of the most number of negative comments.

Table 1: Candidate feedback on FCE for Schools parameters in percentages

Part of paper		Agree that	
		Time enough	Words enough
Part 1	Initial trial	88.3	67.3*
	Final trial	91.1	77.5*
Part 2	Initial trial	91.4	83.1
	Final trial	92.3	79.9

* statistically significant difference

Table 2 shows the percentage of candidates producing a particular number of words or less. Interestingly, the proportion of candidates who produced responses in 190 words or less corresponds closely with the proportion of candidates who agreed that the word count guidelines were sufficient, as shown in Table 1. It can also be seen that only a small proportion of candidates produce responses of less than 160 words.

Table 2: Percentage of candidates producing x words or less

Words	Part 1	Part 2
120	1	3
130	4	10
140	12	20
150	21	33
160	29	48
170	50	67
180	67	77
190	77	86
200	85	92
210	92	95
220	95	98

Should the upper limit of the word count range be adjusted upwards some more (since not everyone agrees that the limit is appropriate and since a number of candidates produce more words)? Part 2 can serve as a useful barometer for answering this question, as it consists of tasks from the

previous edition of FCE that remain unchanged. In spite of the upper limit being increased from 170 to 190 between the first and final trials, the proportion of candidates agreeing with the word count being sufficient has not changed (Table 1). This indicates that, for this task about which we know more, the limit is about right. Presumably, the roughly 20% who do not agree are stronger candidates who want more scope to demonstrate their strong B2 abilities. Looking now at Part 1 – the new essay task – it can be seen in Table 1 that the proportion of candidates agreeing that the word count is sufficient is now close to the 80% benchmark set by Part 2. This provides support to the notion that an appropriate word limit had been arrived at.

In support of this, when the responses were marked, trial data showed that the number of words produced by candidates at the just passing level is 179 and 183 for Parts 1 and 2, respectively.

A case could indeed be made to increase the guide range some more. Militating against this is the observation that as the recommended word count increased, so did candidate production – suggesting that there is potentially no end to this upward revision.

Putting together all the evidence, it can be concluded that around 180 words are needed to satisfy the task requirements. At that point, the vast majority of candidates feel like they have had sufficient opportunity to demonstrate their ability, and examiners agree that responses of such length would indeed meet the standard for CEFR Level B2. The suggested word range allowing fewer and more words than that helps to make the test accessible to weaker and stronger candidates. Test users are also reminded that the range given merely serves as a guide, and candidates are free to produce more if they feel this is necessary and relevant to the task.

For CAE, the same process was followed. Initial specifications of 75 minutes to produce two samples of 200–240 words in the end became 90 minutes to produce two samples of 220–260 words.

Do the tasks engage appropriate cognitive processes?

As exams like CAE are increasingly used for purposes of entry to Higher Education, it is important that test tasks engage some of the same cognitive processes that learners will employ in that context.

Drawing from various theorists (Field 2004, Hayes and Flower 1980, Kellogg 1994, 1996), Shaw and Weir (2007) identify six levels of cognitive processing in writing: macro planning, organisation, micro planning, translation, monitoring, and revising. In addition, they discuss Scardamalia and Bereiter's (1987) characterisation of writing as being either knowledge telling or knowledge transforming. The former merely involves task execution, where knowledge is simply brought in and displayed on the page, whereas the latter involves problem-solving, where arbitration is necessary and knowledge is discovered in the process of writing. Writing in the university context clearly involves all six levels of cognitive processing, and requires both knowledge-telling and knowledge-transforming operations.

Based on the essay task alone, an argument can be made that the six levels of processing are covered, as well as the employment of knowledge telling and transforming modes. Having to select two of three concepts to cover and being able to discuss their merits relative to each other requires *macro planning* and *organisation*. As candidates write, *micro planning* and *translation* (of ideas into linguistic form) are necessarily involved. Trial candidate responses show strikethroughs, erasures and other markings which provide evidence of *monitoring* and *revision* (Elliott, Lim and Galaczi 2012, Lim 2013). Discussion of chosen concepts involves some knowledge telling, and arguing for one concept over another is likely to require knowledge-transforming writing. The second task gives candidates a choice of producing a letter, a proposal, a report, or a review, and these tasks have many of the same characteristics as the first task in terms of cognitive processing.

It needs to be stated that whether writing is knowledge telling or transforming ultimately resides in the writer. However, tasks can, and in this case do, encourage candidates to involve transforming in addition to telling. The marking criteria (Lim 2012) also do this. To obtain a passing mark on CAE, a writer will need to have used 'the conventions of the communicative task effectively to hold the target reader's attention and communicate straightforward and complex ideas'. Organisational patterns are needed to produce well-organised and coherent texts, and complex forms need to be used with control and flexibility. Successful achievement of these, given the task, will require knowledge transformation, and candidates who are not able to do so will not be successful in the test.

In view of the above, a strong argument can be made that the Writing paper does sample the cognitive processes involved in university writing, successful completion of which provides evidence of capability to engage in the same.

How do the new tests compare to the old tests?

The peculiar challenge of revising tests is that they need to become better (or else what's the point of revising) but that they should also stay the same (or else they would not be the same test). That is, the test itself must represent an improvement in some way over the old version, but the same standard should be maintained, especially in view of FCE, FCE for Schools and CAE being tests tied to particular external frameworks and standards.

As previously explained, the tasks were designed with the CEFR in mind, and candidates' responses to these tasks are marked using assessment scales that were themselves developed with reference to the CEFR and empirically validated to be at the right levels (Lim 2012). That being the case, there is *prima facie* a strong case for the tests being at the intended levels.

When teachers were queried (Table 3), they agreed that the new versions of the FCE and FCE for Schools tests indeed represented an improvement over the old versions, the most prevalent reason given being that the compulsory essay offers a greater opportunity to display range and is a useful and relevant function for candidates to become skilled in. But

Table 3: Teacher opinions on comparability, FCE for Schools

Opinion on quality	**%**
Better	46
Same	35
Worse	19
Opinion on difficulty	**%**
Easier	12
Same	61
Harder	27

equally importantly, teachers also felt that the same standards were maintained in the new tests.

Whether the new task is easier, harder or the same can be investigated by comparing it to performance on the old task. Table 4 shows the scores obtained by candidates on each of the criteria during the initial trial for FCE and FCE for Schools. The Part 1s are the new essay tasks, and the Part 2s the task from the old version of the tests. It can be seen that marks are comparable across criteria, the lone exception being the Content mark for the FCE essay task then being trialled. (This provided insight into the ways in which the prototype task was working well or not, and the specifications and item writer guidelines were improved as a result of this.) The disparity in Content marks disappeared in subsequent trials, something that is also borne out by candidate performance since the new test went live.

Table 4: Average scores in the initial trial

Exam and part		**Content**	**Communicative Achievement**	**Organisation**	**Language**
FCE	Part 1	3.8	3.0	2.8	2.5
	Part 2	4.2	3.2	2.7	2.7
FCEfS*	Part 1	4.0	3.6	3.3	3.1
	Part 2	4.0	3.5	3.2	3.0

* FCE for Schools

What effect might the changes have on teaching and learning?

In developing tests, it is vital for test providers to account for the effects these tests might have on stakeholders. It is of course the case that the impact of a test cannot be fully determined until it has been in use for a period of time. That notwithstanding, in keeping with the Cambridge English approach of impact by design (Saville 2009, 2012), the opinions of candidates, teachers and examiners was solicited to assess the potential impact of the new Writing papers.

In the first round of trialling, 9.1% of candidates expressed a negative opinion of the new CAE Writing test. By the final trial, the proportion of candidates with a negative opinion had decreased to 3.5%. For their part, teachers and examiners were positive about the changes. This can also be seen in the potential influence of the changes on their teaching. Teachers indicated that they would not simply give their students more essay writing practice. Rather:

> More [classroom] time would be spent on debates, encouraging opinion which students could then use in an essay.
>
> Focus on planning and editing strategies to enable candidates to complete both tasks.
>
> More focus on differentiating between the different text types.

Teachers are suggesting that in teaching their students to adequately deal with the essay task they will focus on increasing range and function awareness as well as organisational and argumentative skills, which will all be of benefit beyond just the CAE Writing paper. The potential for the changes to have positive washback into teaching is therefore quite high.

Conclusion

The revisions to FCE, FCE for Schools and CAE provided an opportunity to re-imagine what the essay tasks might look like within the Writing papers of these Cambridge English exams. As this article has hopefully made apparent, the process was quite complex and involved. Many factors had to be considered and decisions made in order to produce valid and useful tests. The result of all these are the same Cambridge English exams as before, only better.

References

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

Elliott, M, Lim, G S and Galaczi, E D (2012) *FCE and CAE construct validation study (Part 2)*, Cambridge: Cambridge ESOL internal report.

Field, J (2004) *Psycholinguistics: the Key Concepts*, London: Routledge.

Hayes, J R and Flower, L S (1980) Identifying the organisation of writing processes, in Gregg, L W and Steinberg, E R (Eds) *Cognitive Processes in Writing*, Mahwah: Lawrence Erlbaum, 1–28.

Kellogg, R T (1994) *The Psychology of Writing*, Oxford: Oxford University Press.

Kellogg, R T (1996) A model of working memory in writing, in Levy, C M and Ransdell, S (Eds) *The Science of Writing*, Mahwah: Lawrence Erlbaum, 57–72.

Lim, G S (2012) Developing and validating a mark scheme for Writing, *Research Notes* 49, 6–10.

Lim, G S (2013) *First and Advanced revised writing tasks trial 3*, Cambridge: Cambridge English internal report.

Saville, N (2009) *Developing a model for investigating the impact of language assessment within educational contexts by a public examination provider*, unpublished thesis, University of Bedfordshire.

Saville, N (2012) Applying a model for investigating the impact of language assessment within educational contexts: The Cambridge ESOL approach, *Research Notes* 50, 4–8.

Scardamalia, M and Bereiter, C (1987) Knowledge telling and knowledge transforming in written composition, in Rosenberg, S (Ed) *Advances in Applied Psycholinguistics, Volume 2: Reading, Writing and Language Learning*, Cambridge: Cambridge University Press, 142–175.

Shaw, S D and Weir, C J (2007) *Examining Writing: Research and Practice in Assessing Second Language Writing*, Studies in Language Testing volume 26, Cambridge: UCLES/Cambridge University Press.

UCLES (2015) *Cambridge English: Advanced Handbook for Teachers*, Cambridge: UCLES.

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

Appendix
Sample CAE essay task

You **must** answer this question. Write your answer in **220 – 260** words in an appropriate style on the separate answer sheet.

---

1 Your class has attended a panel discussion on facilities which should receive money from local authorities. You have made the notes below:

Which facilities should receive money from local authorities?

- museums
- sports centres
- public gardens

Some opinions expressed in the discussion:

"Museums aren't popular with everybody!"

"Sports centres mean healthier people."

"A town needs green spaces – parks are great for everybody."

Write an essay discussing **two** of the facilities in your notes. You should **explain which facility it is more important** for local authorities to give money to, **giving reasons** in support of your answer.

You may, if you wish, make use of the opinions expressed in the discussion, but you should use your own words as far as possible.

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

Studies in Language Testing

An indispensable resource for anyone interested in new developments and research in language testing

To find out more about our full list of publications:

www.cambridge.org/elt/silt

www.cambridgeenglish.org/silt

9334017779

To subscribe to *Research Notes* and download previous issues, please visit:
www.cambridgeenglish.org/research-notes

Contents:

For further information visit the website:
www.cambridgeenglish.org

Cambridge English
Language Assessment
1 Hills Road
Cambridge
CB1 2EU
United Kingdom

www.cambridgeenglish.org/helpdesk

THE QUEEN'S AWARDS FOR ENTERPRISE: 2015

ALTE Association of Language Testers in Europe

A DIVISION OF CAMBRIDGE ASSESSMENT

All details are correct at the time of gong to print in November 2015