UNIVERSITY of CAMBRIDGE
ESOL Examinations

Research Notes

Issue 49

August 2012

ISSN 1756-509X

UNIVERSITY *of* CAMBRIDGE
ESOL Examinations

Research Notes

Issue 49 / August 2012
A quarterly publication reporting on research, test development and validation

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

Research Notes

Contents

Editorial notes

Welcome to issue 49 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within University of Cambridge ESOL Examinations (Cambridge ESOL). The theme of this issue is performance testing, largely focusing on written assessment.

In the first article, Helen Spillett outlines the revision of the *Cambridge English: Proficiency (CPE)* Writing paper, part of the wider revision of the whole examination. Next, Gad S Lim outlines the development and validation of a mark scheme for General English and Business English exams, a process which took almost two years. Angela ffrench, Graeme Bridges and Joanna Beresford-Knox share with the reader the review and subsequent revision of the quality assurance process for Cambridge ESOL's 650 Writing examiners.

The next pair of articles is based on Master's dissertations by our staff. Graham Seed explores perceptions of authenticity in test tasks designed to test academic writing in relation to Cambridge English exams, while Heidi Endres reports on her comparability study of computer-based and paper-based *Cambridge English: Preliminary (PET)* Writing tests.

Lucy Chambers, Evelina Galaczi and Sue Gilbert then explore whether test takers' familiarity with each other affects their Speaking test performance in a paired *Cambridge English: First (FCE)* Speaking test.

We are also pleased to publish a paper by Helen Donaghue and Jason Thompson of the Higher Colleges of Technology in the United Arab Emirates (UAE). Donaghue and Thompson's paper discusses the implementation of Khalifa and Weir's 2009 socio-cognitive model of reading within a UAE context.

This paper is followed by a report on the latest ALTE activities and forthcoming events.
We end this issue by inviting you to complete a short reader survey, also available online at http://www.surveymonkey.com/s/Research_Notes_Readership_Survey

The revision of the Cambridge English: Proficiency Writing paper

HELEN SPILLETT ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ESOL

Introduction

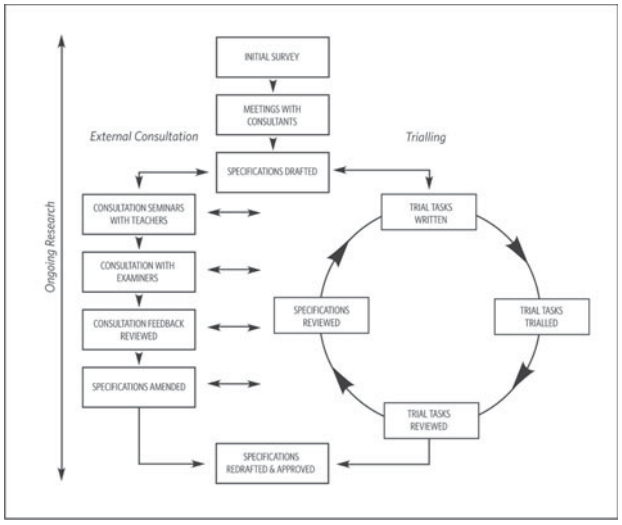
This article outlines the revision of the *Cambridge English: Proficiency* (*CPE*) Writing paper which has taken place as part of the wider revision of the *Cambridge English: Proficiency* examination as a whole.[1]

Background to the current *Cambridge English: Proficiency* revision

Cambridge English: Proficiency, formerly known as the *Certificate of Proficiency in English*, was first launched in 1913. It was aimed at those wanting to use English to teach at university level or to enter the diplomatic services. Nearly 100 years later, *Cambridge English: Proficiency* still provides the most in-depth and thorough examination for those who need high-level English to excel. Teaching and testing methodology and practices have evolved since 1913 and, to reflect these changes, *Cambridge English: Proficiency* was revised significantly in 1945, 1967, 1984 and 2002. The most recent revision process began in 2009 and the launch of the new examination will coincide with *Cambridge English: Proficiency*'s centenary in 2013.

All Cambridge English examinations are periodically updated to reflect the latest research and developments in teaching and assessment, and to ensure that they remain appropriate for candidate needs. The revision cycle includes consultation with teachers, centres and other stakeholders (Figure 1).

Figure 1: Model of the revision process



As part of this process Cambridge ESOL consulted 700 teachers and 170 schools and examination centres in November 2009. In general, analysis showed that both groups felt that:

- content in the revised *Cambridge English: Proficiency* should be suitable for general, study and career advancement purposes
- the examination should be shorter (many felt that some individual papers contained too many tasks)
- a computer-based version would be popular
- the testing focuses and coverage in the current examination were about right.

Teachers and centres also said that while enhanced job prospects and recognition in the higher education sector were important motivators for candidates taking the examination, some potential candidates were put off taking *Cambridge English: Proficiency* because they perceived it as being 'too difficult' and were afraid of failing.

This research, combined with additional work carried out by Cambridge ESOL's specialist test development and research teams, working closely with external consultants and senior examiners, led Cambridge ESOL to frame a review that would seek to make *Cambridge English: Proficiency*:

- more suitable for Higher Education study purposes
- more suitable for career advancement purposes
- look and feel fresher and more modern
- more appealing to the target age group
- shorter than the current examination, while retaining current testing focuses and maintaining the difficulty level
- compatible with computer-based testing.

In considering specific changes, the following questions were also borne in mind:

1. Were there any overlaps in testing focus in the current examination that could be removed to make the examination more efficient but no less comprehensive in its testing coverage?
2. Could any tasks be better situated elsewhere, either within the same paper or in another paper (see below for the papers included in the revised and the current *Cambridge English: Proficiency* examination)?
3. Could the format of the examination conform more closely to the evolving standardised format of other Cambridge English exams?

[1] This article draws on information presented in the Revision Bulletins (Cambridge ESOL 2010, 2011a, 2011b, 2012a, 2012b).

Table 1: Comparison of revised and current Cambridge English: Proficiency examinations

Revised Cambridge English: Proficiency		**Current Cambridge English: Proficiency**	
Paper/timing	**Content**	**Paper/timing**	**Content**
Reading and Use of English 1 hr 30 min	7 parts 53 questions	**Reading** 1 hr 30 min	4 parts 40 questions
		Use of English 1 hr 30 min	5 parts 44 questions
Writing 1 hr 30 min	2 parts: one compulsory question; one from a choice of 5 (including the set text options)	**Writing** 2 hours	2 parts: one compulsory question; one from a choice of 5 (including the set text options)
Listening 40 min (approx.)	4 parts 30 questions	**Listening** 40 min (approx.)	4 parts 28 questions
Speaking 16 min (approx.)	3 parts: interview; collaborative task; individual long turns and follow-up discussion	**Speaking** 19 min (approx.)	3 parts: interview; collaborative task; individual long turns and follow-up discussion
4 papers **Total timing: 3 hours 56 minutes**		**5 papers** **Total timing: 5 hours 59 minutes**	

The revised *Cambridge English: Proficiency* examination

The main changes to the examination overall can be seen in Table 1 above, which compares the revised and the current test specifications.

In brief, the main changes are:

- The five-paper format has been changed to a four-paper format, with the Reading and Use of English papers merged into one paper. This addresses issues of overlap in testing focus and produces a more efficient test without reducing coverage.[2]
- The overall length of the examination has been reduced by 34% in total time. This facilitates more efficient test administration and helps to address the issue of candidate fatigue, while at the same time maintaining the essential qualities of validity and reliability.
- The summary writing task in the Use of English paper has been removed but a summary element is integral to the new Part 1 Writing task (see Table 2). This retains a testing focus which has been identified as a key indicator of C2 level performance in a more prominent form.

Specific issues for Writing

As can be seen in Table 1, the current Writing paper consists of a compulsory question in Part 1 and a choice of one question from five in Part 2. In Part 1, candidates are asked to write an article, an essay, a letter or a proposal in response to instructions and a short text or texts. All questions in this part have discursiveness as their main focus. For example, candidates may be required to defend or attack a particular argument or opinion, compare or contrast aspects of an argument, explain a problem and suggest a solution, or make recommendations having evaluated an idea. In Part 2, candidates choose one from four tasks, one of which offers a set text option. In the second part, candidates are able to select the task and topic which best suits their interests or which they think they can perform best on. The task types available in Part 2 are: article, letter, report, essay (set text questions only) and review, report, and proposal (non-set text questions only). The word range for both questions is 300–350 words.

In the context of the overall *Cambridge English: Proficiency* revision the opportunity was taken to examine the testing focus of the current Writing paper in light of the Common European Framework of Reference (CEFR) descriptors for C2 level writing proficiency (Council of Europe 2001). According to these descriptors, a proficient writer can produce fluent coherent prose in a variety of genres, and:

> can write clear, smoothly flowing text in an appropriate style . . . can write complex letters, reports or articles which present a case with an effective logical structure which helps the recipient to notice and remember significant points . . . can write summaries and reviews of professional or literary works (Council of Europe 2001:27).

The ability to produce fluent and coherent letters, reports, articles and reviews was already well covered in both parts of the current test, with varying output text types for the compulsory Part 1 question (with the constant of a discursive focus), and a choice of output text types in Part 2 including the option of writing about set texts. With regard to the writing of summaries, a further CEFR descriptor states that the writer should be able to 'summarise information from different sources, reconstructing arguments and accounts in a coherent presentation of the overall result' (Council of Europe 2001:96). An objective of the revision process was to ensure coverage of this further aspect of C2 writing proficiency (reading-into-writing and synthesis skills) in the revised *Cambridge English: Proficiency* examination.

As part of the overall aims of the *Cambridge English: Proficiency* revision, any changes to the Writing paper were also to be considered in the context of a shorter examination. This aim to reduce timing, which came out of the consultation process, contributed to the decision to combine the Reading and Use of English papers. As a part of this reduction from five papers to four, the possibility of moving the summary task, which is Part 5 of the current Use of English paper, to

[2] For further discussion of the rationale for combining the Reading and Use of English papers, see, *Revision Bulletin No. 3*, October 2011, www.cambridgeesol.org/assets/pdf/exams/cpe/cpe-revision-bulletin-3.pdf

the revised Writing paper was considered. A clear argument for doing this could be made as the summary task tests both reading and writing skills. However, the eventual specifications for a writing task including a summary element for the revised Writing paper evolved and developed as part of the revision process of consultation, trialling and analysis.

Accordingly, these specific issues were of key importance in the revision process for the *Cambridge English: Proficiency* Writing paper:

- How should the Part 1 and Part 2 tasks be modified to provide comprehensive assessment coverage of the identified components of C2 writing proficiency, including a summary element?
- Which combination of task types should be retained?
- What would the optimum timing of the revised Writing paper be to meet the demands of the revision whilst giving candidates adequate opportunity to demonstrate C2 proficiency?
- What should the impact be on recommended word ranges for candidate output?

The Writing revision process and outcomes

After the first stage of consultation with stakeholders, including external consultants, examination centres, teachers and examiners, two different scenarios were proposed for the revised Writing paper. For both, it was thought desirable to retain the optional text-type Part 2 format, including the questions on set texts, but to reduce the required candidate output length for each task to 250 words (from 300–350 words in the current examination) and to craft tasks with this revised output length in mind. For Part 1, two alternative revised task types were considered: a 250-word compulsory essay or a 150-word summary task. The Research and Validation department at Cambridge ESOL conducted an empirical and theoretical review of VRIP implications of these proposals (i.e. effect on validity, reliability, impact and practicality) and recommended the following format of the new Part 1, to be used in a sample test in the first round of trialling:

Table 2: The new Part 1 of the Writing paper

Candidates to:	Reason
• Read two short input texts. • Write a summary and evaluation, integrated in a coherent essay.	• for proper construct coverage and alignment to the CEFR: writing construct at C2 level requires summary and evaluation skills from different sources in addition to ability to write continuous prose. • for positive washback, to ensure that these writing skills are taught. • to include skills relevant in academic contexts.
• Produce at least 250 words.	• to give scope for candidates to integrate both summary and evaluation in a continuous piece of writing. • for reliability – markers would have more to evaluate than in a 150-word summary task.
• Have 1.5 hours for Parts 1 and 2 combined.	• to give adequate time for task fulfilment. • this would be in line with timing for other *Cambridge English: First* and *Cambridge English: Advanced* Writing papers.

In all there were four rounds of trialling, with participation from authorised examination centres in key *Cambridge English: Proficiency* candidature areas in the UK, Europe, South America and Asia. Trial tests contained different combinations of sample Part 1 and Part 2 questions and cross-rating by a small team of senior examiners. As a result of feedback from candidates, teachers and examiners and analysis of results, proposed task specifications, rubrics and recommended word-length ranges were progressively refined.

The minimum required word count for the new *Cambridge English: Proficiency* Writing papers (240 words for Part 1; 280 words for Part 2) was arrived at as a result of this trialling and analysis. Initially, it was suggested that the minimum be 250 words for both tasks, as this figure is more in line with other Cambridge English: Writing papers, and trialling proceeded on this basis. However, the following outcomes were observed: (1) candidates on average actually wrote more than 280 words; and (2) where Part 2 was concerned, responses that were deemed underdeveloped were all less than 280 words. Together these provided evidence that, in general, a higher minimum is required to produce an adequate response to Part 2. Reducing the required number of words too far would be likely to lead to more underdeveloped responses and, when penalised for that, lead to lower scores for candidates and affect the discrimination and reliability of the paper. The final word length ranges (240–280 for Part 1; 280–320 for Part 2, average overall expected word length 560) were agreed. This also took into account the reduced timing for the Writing paper and the time needed in Part 1 for reading and identifying key points in the input texts. The evidence from trial candidates' responses and feedback indicates that the allotted time of 90 minutes is sufficient for completing a paper with a combined minimum of 520 words. Stronger candidates are also able to produce longer responses, with greater opportunity to demonstrate level, in the same time frame. In this connection it should be noted that candidates are not penalised for over-length scripts *per se* under the new assessment scale for writing (see Lim's article in this issue) and so exceeding the recommended word range is acceptable.

The agreed wording of the rubric for the new Part 1 task reflects the feedback from trialling on the requisite level of clarity and simplicity, enabling candidates to focus fully on the input texts themselves. The rubric reads:

> *Read the two texts below.*
>
> *Write an essay summarising and evaluating the key points from both texts. Use your own words throughout as far as possible, and include your own ideas in your answers.*

This highlights these features of the new task: there are two input texts; each contains clear main points; candidates must identify and integrate a summary of these points, and their own views on the topic in a coherent essay. Although the new Part 1 is clearly to some extent an integrated skills task, the testing aim is to focus on writing; accordingly, the input texts are short (approximately 100 words each), at a level below C2, and clearly express no more than two key points. The idea is not for candidates to produce a discrete summary followed by a brief essay but rather to incorporate the summary

coherently into the argument presented. The input texts are chosen to encourage an abstract discussion eliciting C2-level language, triggered by a concrete context.

Quantitative and qualitative evidence from trialling during the review process indicates that the old and new tests are comparable in difficulty. Marks obtained by candidates who responded to both old and new tasks were not statistically different, and pass–fail decision consistency on old and new tasks was extremely high (approximately 90%). When queried, candidates and examiners overwhelmingly thought that the new test is preferable to the current one while remaining at the same level of difficulty. In addition, the new writing assessment scale (see Lim's article in this issue) will be in use when the new Writing paper is launched in March 2013. This new analytical scale is explicitly linked to Level C2 descriptors on the CEFR, so outcomes should be very robust in terms of indicating whether a candidate's output meets these C2 criteria or not. The evidence from trialling indicates that after training, examiners can mark the new Writing paper to an acceptable level of reliability.

Production of the new task for inclusion in live tests has now begun and early indications from trialling of live versions of the new task with candidates preparing for the current *Cambridge English: Proficiency* examination have been broadly positive. Evidence shows that candidates at C2 level are able to identify key points and develop these in conjunction with their own ideas in an essay with a discursive focus. Moreover, the word range of 240–280 words provided sufficient scope for C2-level candidates to respond well to the new Part 1 task. Again, the strongest candidates consistently produced responses of more than the 300 words. Simultaneously, support materials containing sample tasks and rated scripts are being prepared to be published in time for teachers and candidates to prepare for the revised paper before its launch in March 2013.

The new Part 2 will no longer contain the option of the proposal text type. Qualitative feedback from teachers and examiners on task performance has indicated that the parameters of the proposal text type (not one included in the list from the CEFR quoted above) were the least clearly understood by the global candidature. Many of the language functions involved in writing a proposal are generally similar to those involved in the writing of reports so it was decided that the exclusion of the proposal as a text type should not have an effect on construct coverage of the revised test.

The retention of the set texts option addresses the provision of modern and contemporary literature as a means to engaging with language at C2 level. It is commonly the case that the questions which elicit the highest mean score on a *Cambridge English: Proficiency* Writing paper are the set text questions and it is notable that many students and teachers testify to the benefits gained by studying a set text, even if that option is not taken up in the actual examination (Fried-Booth 2004). Whether in the classroom or by studying privately, there continue to be advantages, both tangible and intangible, which result from being able to broaden cultural and linguistic horizons (ibid).

Conclusion

The changes to the *Cambridge English: Proficiency* Writing paper are summarised in Table 3 below.

Table 3: Comparison of revised and current Cambridge English: Proficiency Writing papers

	Revised Writing paper	Current Writing paper
Timing	1hr 30 min	2 hours
Number of parts	2	2
Part 1	Compulsory question: essay involving summary and evaluation of key points from two input texts	Compulsory question: varied text type
Part 2	One question from a choice of 5, including set texts; text types: article, review, report, letter, essay (set texts only)	One question from a choice of 5, including set texts; text types: article, review, report, letter, proposal, essay (set texts only)
Word ranges	Part 1: 240–280 Part 2: 280–320	Part 1: 300–350 Part 2: 300–350
Word range total	520–600	600–700

The new *Cambridge English: Proficiency* Writing paper ensures a consistency of approach in maintaining comprehensive coverage of writing skills at C2 level, while simultaneously reducing the timing of the paper and length of the required output. Although the reading input for Part 1 has been increased, candidates are required to write fewer words. Furthermore, the shift in Part 1 designed to synthesise reading, summary and writing skills provides the opportunity to prepare more effectively for university level study and professional career development.

References and further reading

Cambridge ESOL (2010) *Revision Bulletin No 1*, available online: www.cambridgeesol.org/assets/pdf/exams/cpe/

Cambridge ESOL (2011a) *Revision Bulletin No 2*, available online: www.cambridgeesol.org/assets/pdf/exams/cpe/

Cambridge ESOL (2011b) *Revision Bulletin No 3*, available online: www.cambridgeesol.org/assets/pdf/exams/cpe/

Cambridge ESOL (2012a) *Revision Bulletin No 4*, available online: www.cambridgeesol.org/assets/pdf/exams/cpe/

Cambridge ESOL (2012b) *Revision Bulletin No 5*, available online: www.cambridgeesol.org/assets/pdf/exams/cpe/

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

Fried-Booth, D (2004) Set texts in CPE Writing, *Research Notes* 18, 12–17.

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.

Developing and validating a mark scheme for Writing

GAD S LIM RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

Introduction

In 2009, as part of its commitment to continuous improvement of its exams, Cambridge ESOL started to revise the writing mark schemes for the *Cambridge English: Preliminary (PET); Cambridge English: First (FCE); Cambridge English: Advanced (CAE); Cambridge English: Proficiency (CPE)* and *Cambridge English: Business Certificates (BEC)* examinations. This was as part of a cyclical test development model (Saville 2003) and in response to a number of identified operational requirements. A process for developing and validating a new mark scheme was therefore set in motion. This process drew upon Cambridge ESOL's experience in developing mark schemes, most recently the revision of the assessment scales for Speaking for the same group of exams (Galaczi, ffrench, Hubbard and Green 2011). From beginning to end the process took the better part of two years.

This article will first outline the development and validation process, to show how the process conforms to current thinking about scale development. The article will then go into more detail about the specific requirements that attended this revision project, illustrating how each requirement is reflected in the resulting assessment scale and mark scheme (assessment scale is used here to refer to the set of performance descriptors covering the ability range, while mark scheme is a super-ordinate term encompassing the assessment scale as well as the various rules and mechanics surrounding its use). It will hopefully be made clear why the mark scheme took on the form and contains the features that it does.

Overview of the development and validation process

Current thinking argues for empirical validation of assessment scales (Council of Europe 2001, Upshur and Turner 1995). That is, assessment scales should not be constructed relying solely on expert judgement, but should involve a range of methods for establishing their validity. The Common European Framework of Reference (CEFR) classified these methods into three – intuitive, qualitative and quantitative – and recommends using a combination of them in 'a complementary and cumulative process' (Council of Europe 2001:207).

That recommendation was adopted when the Speaking scales for the same Cambridge English exams listed above were revised. Galaczi and ffrench (2007) describe an intuitive phase, a qualitative phase, and a quantitative phase. The same division can be used to characterise the present project:

1. Intuitive phase: Internal and external experts produced reports reviewing current literature and practice, made recommendations regarding design principles, and wrote draft descriptors. An internal working group harmonised the various reports and recommendations to produce a first draft of descriptors.
2. Qualitative phase: Internal and external experts participated in scaling exercises, rank ordering the descriptors. Their questions and perceptions of the descriptors were also captured. Analysis of test performances at multiple levels was conducted to evaluate the extent to which descriptors and performances were congruent. Informed by the analysis, subsequent drafts were produced.
3. Quantitative phase: Writing examiners marked performances in multiple marking trials, involving over 20,000 marks, to determine usability and reliability of the mark scheme. Their questions and perceptions were captured via questionnaires. In view of these, the assessment scale and mark scheme were further tweaked.

What must be said about the above characterisation of the process is that, while it makes clear that all three types of scale development methods were used, it also makes the methods appear more distinct from one another than they really were. In reality, the methods overlapped quite a bit. For example, the scaling exercise in the qualitative phase was analysed using multi-facet Rasch, a quantitative methodology. Similarly, while the marking trial aimed to determine reliability, a quantitative matter, the intuitive input of the examiners continued to be solicited and continued to inform the wording of the descriptors and the mechanics of the mark scheme.

The above characterisation also makes the process appear quite linear, where in ways it was also quite cyclical. For example, scaling the descriptors was not a one-off exercise. Rather, the results of one exercise were used to inform another draft, which was then subject to another scaling exercise, and so on. For that matter, while the above would make it appear there were just three stages and three drafts, there were in fact closer to a dozen.

In the end, what resulted was a mark scheme built around an analytic scale with four sub-scales that examiners could use to mark a range of exams reliably. (Prior to any in-depth examiner training, overall reliability for all exams bar one was greater than 0.83.) While the overall scale covers a range of CEFR levels, sections of it are extracted for use with exams at particular levels (see Appendix A).

Details of specific mark scheme requirements

It is now generally accepted that validation of exam instruments must account not just for what is being measured – the construct – but also for the ways in which these instruments will be used (Bachman and Palmer 2010, Messick 1989, Weir 2005). Mark schemes are no exception.

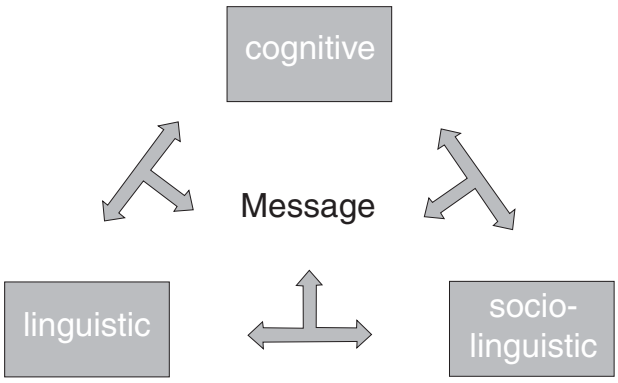
The construct they seek to measure and the ways they are intended to be used in part determine the shape and form they take, which in turn contributes to the argument for their validity. In the following sections the article lays out some of these considerations and requirements, and how these influenced the final product.

Cover the construct

The foremost consideration for any exam is how well it covers the knowledge, skill, or ability it seeks to measure. In this case, the exams that will be using the mark scheme are intended to measure language ability, and in particular writing ability.

At the beginning of the revision process, internal and external experts were commissioned to review the literature on the nature of language ability. The results of that review showed that while scholars differ somewhat in their models' specifics (e.g. Bachman 1990, Canale and Swain 1980, Grabe and Kaplan 1996), they generally agree on two things: (1) they all see language ability in communicative terms as ability for use, and (2) they all see language ability as involving multiple components, which are related to and/or interact with each other. Of components, one can discern three elements: a cognitive element, a language element, and a socio-linguistic element. Writers use these abilities in combination to produce a piece of written text (Figure 1).

Figure 1: Interaction of elements in Writing



A number of assessment scales from other Cambridge English exams and from other test providers were also reviewed to determine the state of the art, so to speak, and they all seemed to reflect the elements identified above. The scales tended to have a criterion (in the case of analytic scales) or descriptor (in the case of holistic scales) that had to do with organisation, sometimes under the headings of coherence and cohesion. It appeared that these were intended to be indirect measures of the cognitive element, how candidates arranged the material to create an effective piece of writing. The scales all accounted for the linguistic element, sometimes dividing vocabulary and grammar into two separate criteria, sometimes not, and in some other cases also had a separate criterion to do with mechanics. The socio-linguistic element was not covered by all the scales, though it appeared that mark schemes for exams which elicited more than one genre of writing were more likely to account for this element. Finally, most scales also accounted for task achievement in some way or other. It appeared therefore that models of writing ability and mark schemes for writing converged with one another.

When the reviewers were commissioned, they were not given specific instructions on whether a holistic or analytic scale was desired. The presumption within the internal working group was the former, the existing scales being holistic. One reviewer, at the outset, proposed an analytic scale whose categories matched the different elements identified by the different models of language ability (Table 1). Another reviewer also proposed a four-criterion analytic scale, but dividing language into separate categories for lexis and grammar while subsuming socio-linguistic aspects such as register under Content. It was ultimately decided that an analytic scale would be best, as this would allow for different aspects of the construct to be more explicitly accounted for, and would in addition bring Writing in line with Speaking, which had previously adopted an analytic approach (Galaczi et al 2011).

As to what the analytic criteria should be, as a result of the various reviews, it became clear that having separate sub-scales for each of the elements of language ability would be best, so as to ensure proper and balanced coverage of the construct. The final scale thus has one criterion each for the cognitive, linguistic, and socio-linguistic elements, plus one criterion for task achievement.

Table 1: Analytic assessment criteria

Reviewer	Final
• Content and Development	• Content
• Communicative Achievement	• Communicative Achievement
• Organisation and Linking of Ideas	• Organisation
• Range and Control	• Language

Create greater coherence across tasks, exams and domains

The mark scheme is intended for use with a range of Cambridge English exams from CEFR Level B1 to C2[1]. In order to create greater coherence between exams, the same assessment scale would be used for exams at the same CEFR level. This meant that the descriptors had to be more general than they otherwise would be, so that they can work across exams.

While sharing some similarities, the exams nevertheless have some differences (see Shaw and Weir 2007 for details). To give an example, within and across exams, the number of points a candidate needs to address in a given writing task differs. The existing mark schemes had complex rules pertaining to task achievement, e.g. if three out of four points are addressed then . . . However, as the new mark scheme has to be useable across exams and across tasks with differing numbers of points, a system based on counting numbers was no longer feasible. In addition, such a system

[1] Cambridge English Writing tasks at CEFR Levels A2 and below are assessed with a content-specific mark scheme.

would not account for the fact that some points are simpler and some more complex, and thus should not be weighted in the same way. The new Content sub-scale reflects a more qualitative approach, phrasing task achievement in terms of the extent to which the intended reader would be informed and whether or not there are omissions, irrelevances, or misinterpretations.

The revision also provided an opportunity to review how other issues (e.g. under and over length responses; varieties of English) should be dealt with. Where length of response is concerned, the writing tasks specify expected range of the output, and candidate responses that had not kept to the guidelines were generally met with automatic penalties. This however did not seem to be in line with a communicative construct, which would emphasise effectiveness of communication. In addition, it led to candidates and examiners spending significant time counting words, which did not seem to be a good use of their time.

The mark scheme has thus removed automatic length-related penalties, allowing for the effects of these to be dealt with under the four sub-scales. Thus, for example, an over length response may result in irrelevance or have an adverse effect on the reader, which might affect its score for Content and Communicative Achievement, respectively. Similarly, an under length response may not exhibit an adequate range of language, and affect its score for Language. In this way, candidates are rewarded and marked down for their performances only in construct relevant ways. If the length of the candidate's response has no effect on the construct as reflected in the assessment scales, then their mark is not affected.

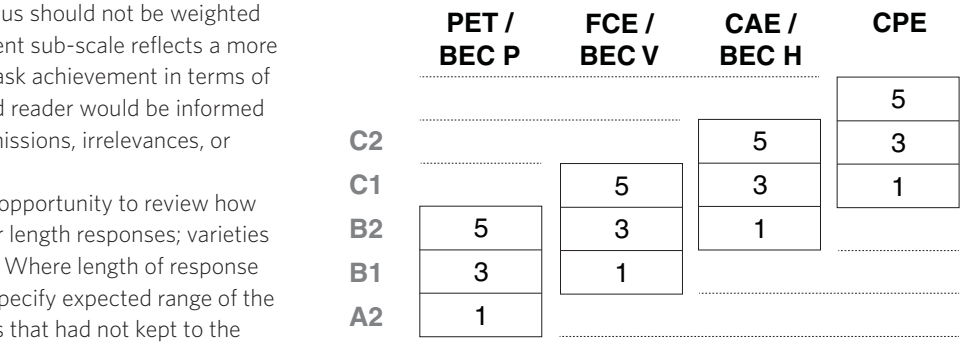
Create explicit links across levels

Another requirement for the mark scheme was creating greater coherence across levels. This converged with Cambridge ESOL's decision to offer enhanced certification. That is, if a candidate taking a B1 level test (say, *Cambridge English: Preliminary*) performs well enough to demonstrate B2 level ability, then they would be given credit for that. If they do not perform at B1 level, they may nevertheless have demonstrated ability to A2 level, and would similarly be recognised in that way.

The basis for these decisions where Reading, Listening and Use of English are concerned is Cambridge ESOL's item banking system, where test items from different exams are calibrated to a common underlying scale, making it possible to know how performance on one exam relates to performance on another exam (Khalifa and Weir 2009, Saville 2003). In order to make the relationship similarly clear for Writing, the assessment scale followed the lead of the recently revised Speaking tests (Galaczi et al 2011) in adopting a 'stacking approach' (Figure 2). So, for example, the descriptor for score point 3 in the B2 scale becomes the descriptor for score point 5 in the B1 scale, and the score point 1 descriptor for the C1 scale. In this way, a relationship across levels is established, making it possible to know how a performance would have been evaluated at a different level.

Achieving a level of course depends partly on the test eliciting performance at that level. This is why at the B1 level, say, the scale only measures up to the B2 level, even if the C1

Figure 2: Relationship of scales at different levels



descriptors could in theory be added on. A B1 task is simply not designed to elicit performance at the C levels, and the B1 scale is capped at B2 for that reason.

Accounts for the CEFR explicitly

Figure 2 showed that the descriptors are related to particular levels of the CEFR. The CEFR levels are based in part on Cambridge ESOL's suite of exams (Hawkey 2009, North 2004, Taylor and Jones 2006) so a relationship already exists between them. However, it was felt that making that relationship explicit was desirable.

In order to do that, an iterative process was followed where (1) descriptors were drafted and revised with particular CEFR levels in mind, covering A2 to C2, with an additional level for performance above the C2 minimum, (2) experts judged which CEFR level each descriptor represented, (3) judgements were analysed using multi-facet Rasch measurement to determine the characteristics of each descriptor, and (4) the findings informed further revision of the descriptors until the experts and the analysis determined that the descriptor was indeed at the intended CEFR level.

These descriptors were all rank ordered as intended within their sub-scale, and many of them were deemed to be at the intended level from the first. This was the case with all the Communicative Achievement descriptors. On the other hand, some descriptors needed to go through several rounds of revision. As an example, Table 2 shows that the language descriptor intended for the C1 level was in the first round of judging deemed to be at a lower level than expected, then at a higher level after revision, before further changes finally resulted in it being judged at the right level.

Table 2: Development of C1 grammar/structure descriptor

Draft	Descriptor	Judged level
1	Uses a range of simple and complex grammatical forms with a good degree of control.	B2
2	Uses a wide range of simple and complex grammatical forms with control and flexibility.	C2
3	Uses a range of simple and complex grammatical forms with control and flexibility.	C1

Other steps were taken to ensure that the CEFR levels were maintained. Writing performances known to be at B2 were analysed for linguistic features to determine the extent to which they matched the descriptors, and the results showed that there was indeed good correspondence between them

(Salamoura and Lim 2012). In addition, as the exams marked using the existing mark schemes were already related to the CEFR, decision consistency has to be observed. That is, performances deemed to be at the level using the old mark scheme should also be deemed to be at the level using the new mark scheme. Results from the marking trial show that a high level of agreement exists between the two mark schemes (Table 3). Thus, while the new mark scheme has a number of features to accommodate new requirements, the levels of the tests are maintained.

Table 3: Decision consistency for exams at different CEFR levels

CEFR level of exam	Decision consistency
B1	93.7%
B2	95.8%
C1	87.5%
C2	90.0%

Consider the consequences

As previously mentioned, accounting for consequential validity is integral to test validity. In relation to this project, it had already been previously decided in the interest of transparency that the assessment scale would be made publicly available. This decision had implications for scale development, which are discussed below.

The literature distinguishes between constructor, assessor, and user oriented scales (Alderson 1991). Depending on who will be using the scales, the way descriptors are phrased would be different. For example, a number of the existing mark schemes use comparative terms and contain a note which says: 'This mark scheme should be interpreted at the *x* level.' This reflects the fact that examiners have many years of experience working with the exams and thus have a clear understanding of the level. On the other hand, teachers and students might not necessarily know what those descriptors mean, making the scale not very useful for them. Thus, the current mark scheme's descriptors needed to be phrased in ways that are concrete, clear and independent (Council of Europe 2001).

In addition, as learners will be able to see the descriptors, it is important that these have a positive effect on their language learning. For this reason, and in keeping with an achievement-oriented rather than deficit-oriented assessment model, the descriptors had to be phrased positively wherever possible (see Council of Europe 2001). This was an important guideline for the working group, and a look at the descriptors would show that the guideline had on the whole been observed. Even the descriptors to do with error emphasise something positive, as in the case of 'while errors are noticeable, meaning can still be determined'. As this example makes apparent, positive phrasing does not mean that a given performance does not exhibit any weaknesses, only that the things that have been done well are emphasised.

Beyond positive phrasing of descriptors, some of the existing mark schemes are careful to account for instances where candidates' errors were due to ambition (e.g. using more complex structures or less common lexical items that they have not yet fully mastered), so as not to penalise them for behaviour that should be encouraged among language learners.

The revised assessment scale accounts for this. At the B1 level, for example, a candidate who decides to play it safe and uses only simple structures will receive at most a 3 for that criterion ('Uses simple grammatical forms with a good degree of control'). On the other hand, the candidate who in addition to simple structures also ventures to use complex forms will necessarily receive a mark no lower than the previous candidate. If they succeed in some cases, then they fit the 5 descriptor which says 'Uses a range of simple and some complex forms with a good degree of control'. If they do not, then their well-controlled simple structures will still earn them a 3.

The concern was also raised during the development process about the possibility of candidates getting marked down more than once for the same thing. That is, the mark scheme might be potentially unfair. But this is not the case. It has already been discussed that the ability being evaluated is multi-componential and, in addition, that the components interact with one another (see Figure 1). Thus, when a candidate's tenuous grasp of the language results in a response that fails to communicate well, it is only right that they get a low mark on both Language and Communicative Achievement. On the other hand, another candidate might be strong linguistically but use entirely the wrong register in their response. In that case, the candidate would receive a good mark for Language but not as good a mark for Communicative Achievement. As can be seen, if candidates are marked down more than once, it is only in those instances where two or more aspects of the construct interacted with each other, and all were affected. Still another candidate might produce a piece of writing that is strong linguistically and communicatively appropriate; in that case, the candidate also deserves their double reward, as it were. What the scale and mark scheme is doing, in other words, is evaluating candidates in construct relevant ways. This makes marking not only fair, but increases its validity as well.

Conclusion

This article has highlighted and described some aspects of the Writing mark scheme development and validation process. It has shown that creating a mark scheme is a complex endeavour, attended by multiple requirements, which inevitably play a role in determining its final shape and form. In this case, the challenge was particularly great, as the mark scheme had to work across tasks, exams, domains and levels. A careful process of decision making was followed throughout, repeatedly going back to the construct as the basis for the mark scheme's features and design. Following an extensive process of development and validation, a product was arrived at which fulfils the requirements of validity, reliability, impact and practicality that all exam instruments must meet.

References and further reading

Alderson, J C (1991) Bands and scores, in Alderson, J C and North, B (Eds), *Language Testing in the 1990s,* London: Macmillan, 71–86.

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

Bachman, L F and Palmer, A S (2010) *Language Assessment in Practice*, Oxford: Oxford University Press.

Canale, M and Swain, M (1980) Theoretical bases of communicative approaches to second-language teaching and testing, *Applied Linguistics* 1 (1), 1–47.

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

Galaczi, E D and ffrench, A (2007) Developing revised assessment scales for Main Suite and BEC Speaking tests, *Research Notes* 30, 28–31.

Galaczi, E D, ffrench, A, Hubbard, C and Green, A (2011) Developing assessment scales for large-scale speaking tests: A multiple-method approach, *Assessment in Education: Principles, Policy & Practice* 18 (3), 217–237.

Grabe, W and Kaplan, R B (1996) *Theory and Practice of Writing*, New York: Longman.

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.

Messick, S (1989) Validity, in Linn, R (Ed.) *Educational Measurement*, New York: Macmillan, 13–103.

North, B (2004) Europe's framework promotes language discussion, not directives, in *Guardian Weekly*, available online: www.guardian.co.uk/education/2004/apr/15/tefl6

Salamoura, A and Lim, G S (2012) *Validating a writing assessment scale: Insights from SLA, corpora, and computational linguistics*, paper presented at the Language Testing Research Colloquium, Princeton, New Jersey.

Saville, N (2003) The process of test development and revision within UCLES EFL, in Weir, C and Milanovic, M (Eds) *Continuity and innovation: Revising the Cambridge Proficiency in English examination 1913–2002*, Studies in Language Testing volume 15, Cambridge: UCLES/Cambridge University Press, 57–120.

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.

Taylor, L and Jones, N (2006) Cambridge ESOL exams and the Common European Framework of Reference (CEFR), *Research Notes* 24, 2–5.

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

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

Appendix

B1 assessment scale[2]

	CONTENT	COMMUNICATIVE ACHIEVEMENT	ORGANISATION	LANGUAGE
5	All content is relevant to the task. Target reader is fully informed.	Uses the conventions of the communicative task to hold the target reader's attention and communicate straightforward ideas.	Text is generally well organised and coherent, using a variety of linking words and cohesive devices.	Uses a range of everyday vocabulary appropriately, with occasional inappropriate use of less common lexis. Uses a range of simple and some complex grammatical forms with a good degree of control. Errors do not impede communication.
4	*Performance shares features of Bands 3 and 5.*			
3	Minor irrelevances and/or omissions may be present. Target reader is on the whole informed.	Uses the conventions of the communicative task in generally appropriate ways to communicate straightforward ideas.	Text is connected and coherent, using basic linking words and a limited number of cohesive devices.	Uses everyday vocabulary generally appropriately, while occasionally overusing certain lexis. Uses simple grammatical forms with a good degree of control. While errors are noticeable, meaning can still be determined.
2	*Performance shares features of Bands 1 and 3.*			
1	Irrelevances and misinterpretation of task may be present. Target reader is minimally informed.	Produces text that communicates simple ideas in simple ways.	Text is connected using basic, high-frequency linking words.	Uses basic vocabulary reasonably appropriately. Uses simple grammatical forms with some degree of control. Errors may impede meaning at times.
0	Content is totally irrelevant. Target reader is not informed.	*Performance below Band 1.*		

B2 assessment scale

	CONTENT	COMMUNICATIVE ACHIEVEMENT	ORGANISATION	LANGUAGE
5	All content is relevant to the task. Target reader is fully informed.	Uses the conventions of the communicative task effectively to hold the target reader's attention and communicate straightforward and complex ideas, as appropriate.	Text is well organised and coherent, using a variety of cohesive devices and organisational patterns to generally good effect.	Uses a range of vocabulary, including less common lexis, appropriately. Uses a range of simple and complex grammatical forms with control and flexibility. Occasional errors may be present but do not impede communication.
4	*Performance shares features of Bands 3 and 5.*			
3	Minor irrelevances and/or omissions may be present. Target reader is on the whole informed.	Uses the conventions of the communicative task to hold the target reader's attention and communicate straightforward ideas.	Text is generally well organised and coherent, using a variety of linking words and cohesive devices.	Uses a range of everyday vocabulary appropriately, with occasional inappropriate use of less common lexis. Uses a range of simple and some complex grammatical forms with a good degree of control. Errors do not impede communication.
2	*Performance shares features of Bands 1 and 3.*			
1	Irrelevances and misinterpretation of task may be present. Target reader is minimally informed.	Uses the conventions of the communicative task in generally appropriate ways to communicate straightforward ideas.	Text is connected and coherent, using basic linking words and a limited number of cohesive devices.	Uses everyday vocabulary generally appropriately, while occasionally overusing certain lexis. Uses simple grammatical forms with a good degree of control. While errors are noticeable, meaning can still be determined.
0	Content is totally irrelevant. Target reader is not informed.	*Performance below Band 1.*		

[2] Writing assessment scales for C1 and C2 can be found in the relevant handbook (www.teachers.cambridgeesol.org/ts/exams)

Quality Assurance: A Cambridge ESOL system for managing Writing examiners

ANGELA FFRENCH ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ESOL
GRAEME BRIDGES ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ESOL
JOANNA BERESFORD-KNOX ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ESOL

Introduction

Cambridge ESOL has a principled approach to the review and revision of all aspects of its examinations, and the Cambridge ESOL model of the test development process (Cambridge ESOL 2011:19) demonstrates the cyclical and iterative nature of this review and revision process, which applies not only to the production of examination papers, but also to operational processes.

In 2009, the Assessment and Operations division (AOG) of Cambridge ESOL embarked on a review of part of its many Quality Assurance (QA) procedures, namely the Examiner Management System for all its Writing papers. This encompassed all stages: from a prospective examiner (PEx) registering an interest in examining on Cambridge English Writing tests, through the different QA checkpoints, to providing feedback on examiner performance.

This paper charts that review process and the subsequent revision of the Writing examiner QA stages. Where 'examiner' is used, this relates to all levels of responsibility:

- Principal Examiner (PE), the most senior examiner who is responsible for the overall standard of the Writing paper
- Team Leader (TL), experienced senior examiners who manage teams of examiners
- Writing examiner (WE).

Background

Cambridge ESOL examiners for Writing and Speaking have always been subject to a rigorous system of QA, from the moment they register an interest in becoming examiners to the point at which monitoring and evaluation of their performance has been fed back to them. This has been well documented (ffrench 2003, Shaw and Weir 2007, Taylor and Galaczi 2011), outlining the system of RITCME: Recruitment, Induction, Training, Co-ordination, Monitoring and Evaluation.

Since 2008, the QA process for Speaking examiners has taken advantage of developments in technology, whereas for Writing examiners, these RITCME stages were dealt with manually and through face-to-face meetings. With the move from a thrice yearly delivery of examinations to an 'on-demand' approach it was acknowledged that a more streamlined approach to Writing examiner management was required.

The review considered how best use could be made of the technologies already available, such as: the Cambridge ESOL Examiner Management System (EMS); the Professional Support Network (PSN) used for Speaking examiners; the online marking tool, scoris® Assessor; Camtasia Studio (screen video capture software) and Adobe Captivate (the electronic learning tool which can be used to author software demonstrations, software simulations, randomised quizzes and score reporting). In this way, the QA process could reach out more effectively to all Writing examiners at different stages of their engagement with Cambridge ESOL.

Seven stages of Writing examiner QA were identified for the new model (Figure 1), the process for which can be seen in Figure 2.

Recruitment and ESOL examiners online

Prospective examiners who are interested in working with Cambridge ESOL apply online via the Cambridge ESOL website careers section. This then takes them to *Cambridge ESOL Online: Examiners* where they can choose what type of examining they would like to do and, in the case of Speaking examiners, where it will take place.

Once the application has been completed and submitted they are given an Application Number. This unique examiner reference ensures Cambridge ESOL can respond quickly about any queries the PEx has regarding their application.

All applications are processed to ensure the applicants meet the Minimum Professional Requirements (MPRs). This includes such things as their professional background, e.g. relevant experience and qualifications. References are then sought by Cambridge ESOL, and successful applicants put on a waiting list. As soon as there is a vacancy, the PEx is invited to complete the Induction stage.

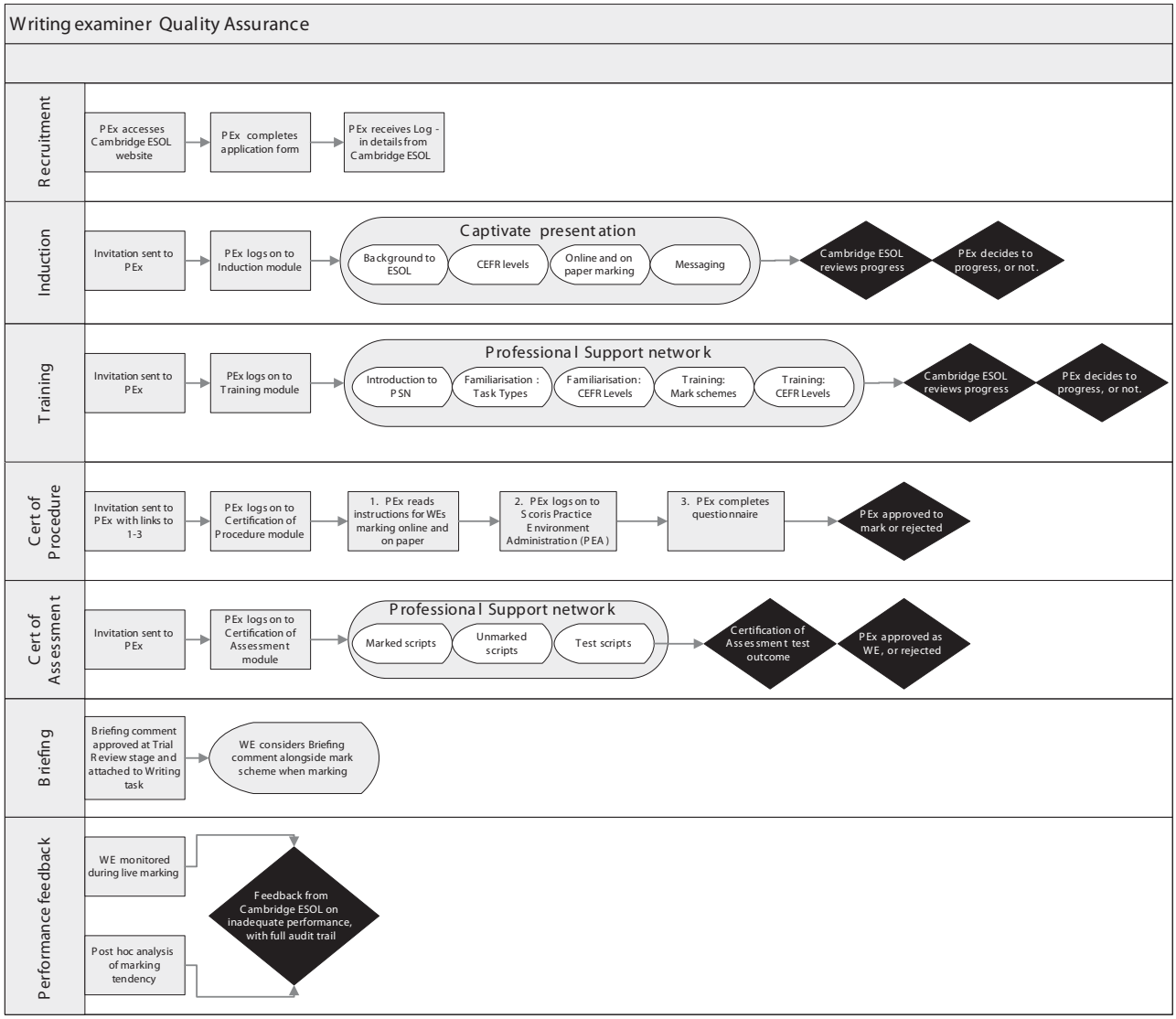
Further use of *Cambridge ESOL Online: Examiners* is described in this paper, in the section entitled 'The examiner experience', at the point in the QA process where examiners begin to engage with live marking.

Induction

Induction is delivered through a 10-minute online presentation using Adobe Captivate. The purpose of the presentation is to give prospective examiners a flavour of what Cambridge ESOL is and does, and what their role as an examiner entails.

Figure 1: Seven stages of Writing examiner QA

Figure 2: Writing examiner QA process



In order to view the presentation, examiners simply need the latest version of Adobe Flash Player and Internet Explorer.

The Induction module gives prospective examiners the opportunity to decide if they want to carry on with the QA process or to withdraw their application, and includes information on:

- the scope of Cambridge ESOL's provision in terms of products and reach
- alignment of the products to the Council of Europe (2001) Common European Framework of Reference for Languages (CEFR)
- the assessment scales used for marking the Writing papers
- the QA process
- marking methods (online and on paper)
- IT specifications required to use the online marking tool scoris® Assessor
- procedures for online marking
- communication and support via the Team Leader system.

Once the Induction module is complete, PExs are then invited to the Training module.

Training

Training is carried out, in the majority of cases, online through PSN, the Professional Support Network. The aims of Training are:

- to familiarise examiners with the domains of Business and General English and the various task types of exams at the CEFR level(s)
- to familiarise examiners with the CEFR-based mark scheme
- to provide practice in applying the mark scheme to scripts at B1, B2, C1 and C2 level through a rank-ordering exercise[1]
- to highlight different procedures for marking, e.g. online and on paper.

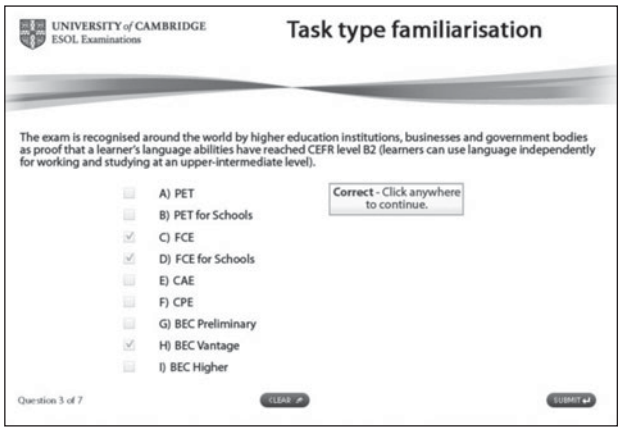
Training is designed to be formative. Quizzes, with the support of relevant documentation (e.g. Teacher's Handbooks, the Cambridge ESOL Mark Scheme for Writing, etc.), are

[1] Writing tasks for Cambridge English examinations at CEFR Levels A1–A2 (i.e. *Cambridge English: Young Learners (YLE), Cambridge English: Key (KET), Cambridge English: Key (KET) for Schools*) are marked by clerical markers (rather than examiners) who are subject to equally rigorous QA measures.

presented to the examiners and immediate feedback is given, including information about each of the questions, how the examiner answered them, and a final score.

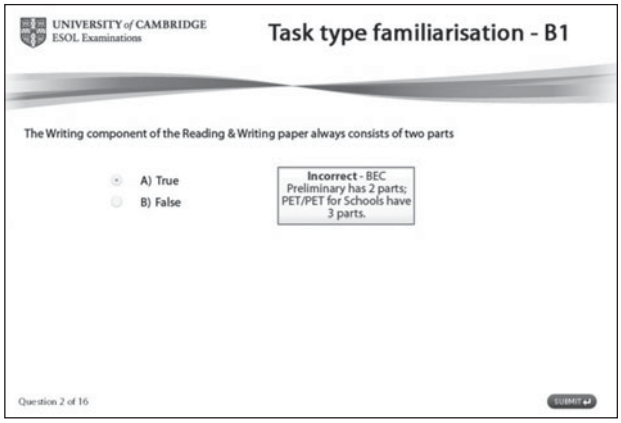
Figure 3 shows an example of a task type familiarisation exercise where examiners are expected to match a series of statements concerning the domains of Business and General English with the appropriate examinations. Feedback is provided instantly.

Figure 3: Task type familiarisation exercise 1



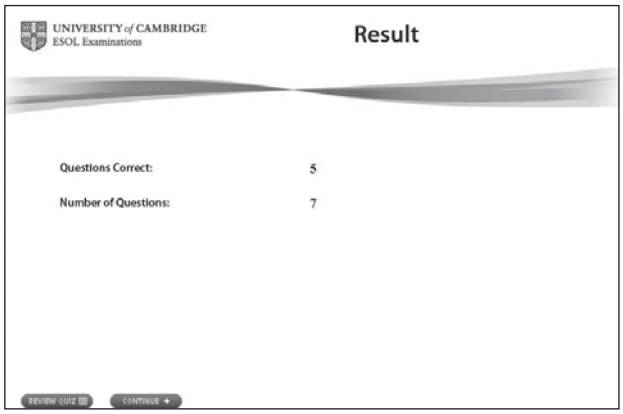
In the exercise shown in Figure 4, examiners answer True/False to various statements on the task types at each level. If an incorrect response is submitted an explanation is provided.

Figure 4: Task type familiarisation exercise 2



A score is provided for each quiz (Figure 5):

Figure 5: Reporting of score



The online content of the module and the software used to deliver it is SCORM-compliant (Sharable Content Object Reference Model). This ensures the examiners go through a fixed set of training material paths and also allows them to bookmark their progress when taking breaks. Examiners can revisit the Training module whenever they wish; this is particularly useful for those who want additional practice prior to a marking session.

The PEx then has to complete two stages of certification before being eligible to mark.

Certification of Procedure

The majority of Cambridge English Writing papers are marked online, but there are occasions when papers are marked on paper. Information about marking on paper is provided in the form of documentation, whereas for online marking, examiners are taken into a practice environment in the online marking system scoris® Assessor and allowed to demonstrate their ability to mark on screen. They then have to complete an online questionnaire that addresses issues in both onscreen and paper marking.

As in Training, Certification of Procedure is designed to be formative. For online marking, examiners are invited to log on to the scoris® Assessor practice environment administration (PEA) tool where they can download a user guide and view an online training guide (Figure 6).

Figure 6: scoris® Assessor online training guide

The guide shows how the examiner would interact with scoris® Assessor: having opened up Section 2 (Figure 7) the examiner just has to click on 'Play' to see how the marking screen works.

Once the examiner has completed the online training tutorials, they log on to scoris® Assessor and practise marking on typical scripts (Figure 8). The marks are not assessed as the purpose of this exercise is to familiarise examiners with using scoris® Assessor.

Having completed the Certification of Procedure module, PExs are then invited to go on to Certification of Assessment. There is no limit to the number of levels an examiner can mark for; they simply need to provide evidence of their marking accuracy.

Figure 7: scoris® Assessor marking screen

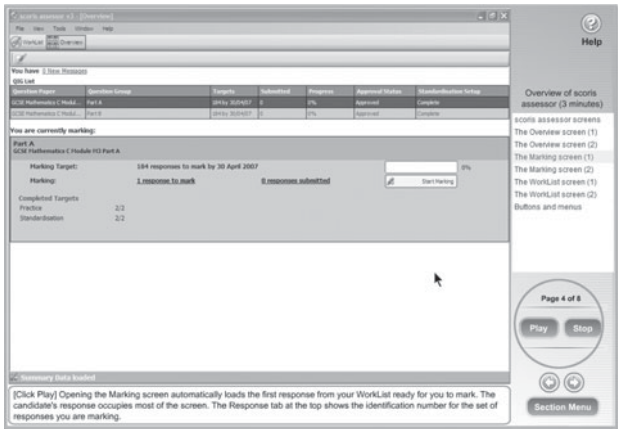
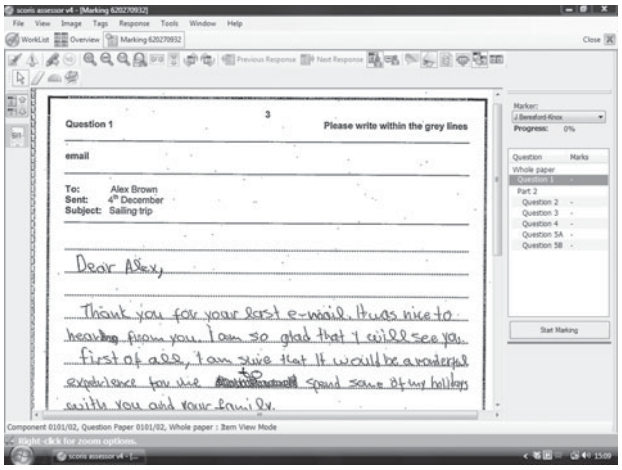


Figure 8: Candidate script



Certification of Assessment

Certification of Assessment is the stage where examiners have to demonstrate that they can mark accurately. This is something that examiners at all levels of responsibility (WEs, TLs, PEs) are required to do, to ensure even the most experienced examiners continue to mark accurately, and is conducted electronically on PSN.

After successful certification at a particular CEFR level, examiners may be asked to mark any of the examinations at that level. For example, at B2 level this may be *Cambridge English: First (FCE)*, *Cambridge English: Business Vantage (BEC Vantage)* or *Cambridge English: First (FCE) for Schools* (Table 1).

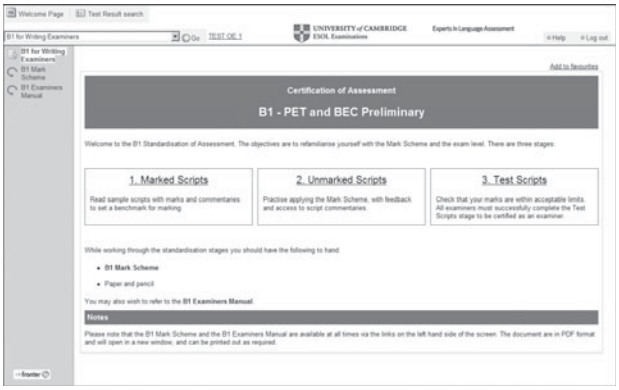
Table 1: Cambridge ESOL exams by CEFR level

CEFR level	General English	Professional English	Exams for Schools
C2	CPE		
C1	CAE	BECH	
B2	FCE	BECV	FCE
B1	PET	BECP	PET

There are three stages to the process (Figure 9):

- marked scripts
- unmarked scripts
- test scripts.

Figure 9: Three stages of Certification of Assessment



Marked scripts

In the first stage, examiners are presented with a script together with the marks awarded for each sub-scale: Content, Communicative Achievement, Organisation, Language (Figure 10) and a commentary to support each mark (Figure 11).

Figure 10: Sample script provided by examiners

Figure 11: Sample script commentaries

This allows them to familiarise themselves with how the language the candidate produces relates to the descriptors for each sub-scale in the assessment scales. The colours that are attached to each sub-scale are also shown on the scripts, highlighting the text which relates to the commentary.

Unmarked scripts

To provide additional practice, examiners are then presented with a fresh set of scripts but are not provided with the marks until they have first attempted to mark them. The examiners then compare the marks they awarded with the official marks and are able to review the commentary supporting each of the official marks.

Test scripts

This is the final stage of the Certification of Assessment process, where examiners are tested on how accurately they assess according to the Cambridge ESOL interpretation of CEFR levels. Tolerance levels are set to establish whether or not an examiner is accurate enough to be invited to mark, and they are given a second chance if they are unsuccessful. Once certificated, examiners can begin to mark.

Briefing

Previously, the Cambridge English mark schemes for some Writing papers comprised two mark schemes: a General Impression Mark Scheme (GIMS) and a Task Specific Mark Scheme (TSMS). While the GIMS focused more generally on aspects of assessment, e.g. effect on target reader, relevant content, organisation, grammar and vocabulary, the TSMS highlighted aspects of the task that needed to be specifically addressed, e.g. nominating a scientist for inclusion in a TV programme, describing their achievements, justifying the nomination, the register appropriate for a letter to the editor of an international magazine etc.

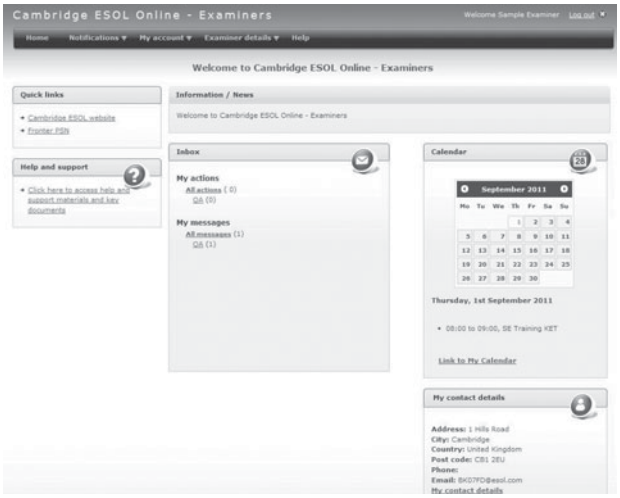
With the introduction of the new mark scheme (see Lim's article in this issue) it was felt appropriate to review this situation. Since the TSMS was not a feature of all exams, and since the presentation of the tasks was completely transparent, it was felt that the TSMS was redundant. See Handbooks for Teachers (Cambridge ESOL 2012) for examples of Writing tasks.

However, it was acknowledged that there might be occasions when some additional information needed to be passed to the examiners. For example, at the trial review stage of the question paper production process, it may transpire that candidates approach a task from different angles, all of which are acceptable. Where information is known in advance of the onset of marking, a Briefing comment is provided to give guidance to the examiner on how to mark the response, clarifying any identified issue. This is available for examiners to download, attached to the question paper.

The examiner experience

Cambridge ESOL Writing examiners take responsibility for managing their marking workload. This is all processed through the *Cambridge ESOL Online: Examiners* portal which enables them to do various tasks to ensure they are prepared, certificated and available for work with Cambridge ESOL. When examiners log on they are presented with the homepage (Figure 12) and from this they can access a variety of different functions, such as messages sent by Cambridge ESOL, QA events, their availability calendar and their contact details.

Figure 12: Cambridge ESOL Online: examiners homepage



Contact details can be kept up to date, and eligibility records can be checked to ensure they reflect the examiner's true status.

The 'Traffic light' system (shown in Figure 13) indicates when the eligibility is due to expire. This allows the examiner to book onto forthcoming events to ensure they are eligible for future marking. As you can see in this example eligibility for Certification of Procedure has expired for all the products the examiner has been marking for, and is about to expire for Certification of Assessment for *Cambridge English: Advanced* (*CAE*) and computer-based (CB) *Cambridge English: Advanced* Writing.

Figure 13: Eligibility status

	Assessment component	Enabled status	Induction	Training	C of P	C of A	Performance feedback	Eligibility status
⊞	CB BEC Vantage Writing	Yes	25/04/2042 WEIM	25/04/2042 WETB2	26/05/2012 WECP	25/04/2013 WECAB2		26/05/2012
⊞	CB FCE for Schools Writing	Yes	25/04/2042 WEIM	25/04/2042 WETB2	26/05/2012 WECP	25/04/2013 WECAB2		26/05/2012
⊞	FCE Writing	Yes	25/04/2042 WEIM	25/04/2042 WETB2	26/05/2012 WECP	25/04/2013 WECAB2		26/05/2012
⊞	CAE Writing	Yes	25/04/2042 WEIM	25/04/2042 WETC1	26/05/2012 WECP	26/06/2012 WECAC1		26/05/2012
⊞	BETS 3 Writing	Yes	25/04/2042 WEIM	25/04/2042 WETB2	26/05/2012 WECP	25/04/2013 WECAB2		26/05/2012
⊞	CB CAE Writing	Yes	25/04/2042 WEIM	25/04/2042 WETC1	26/05/2012 WECP	26/06/2012 WECAC1		26/05/2012

Once they have completed the booking for a particular event, they are sent an invitation which contains all the relevant information regarding the event, such as when and where it is and anything they need to bring to the event, payment, etc. If it is an online event they will be sent a link, if it is face-to-face they will receive details about the meeting (Figure 14).

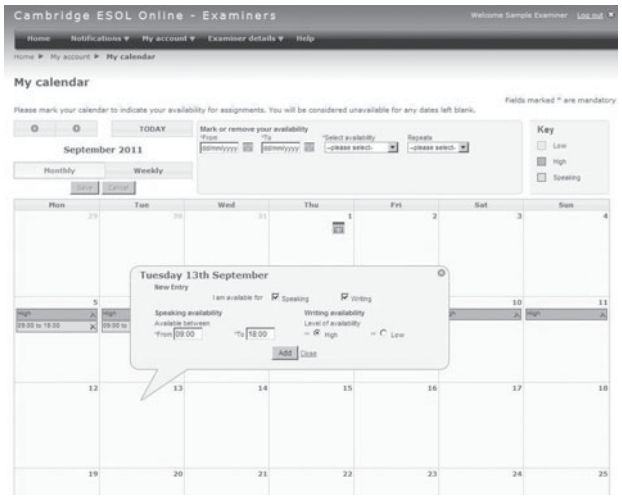
Further to the listed functions the examiners can also view their availability calendar (Figure 15). This allows them to mark up dates for availability and see any QA events that they have been booked onto.

Availability can be added to the calendar for individual days, simply by clicking on the day in question, and filling in the pop-up form, or for a range of dates, by using the 'Mark or remove your availability' menu above the calendar (Figure 16).

Figure 14: Meeting information

Figure 15: Invitation to QA events

Figure 16: Availability calendar

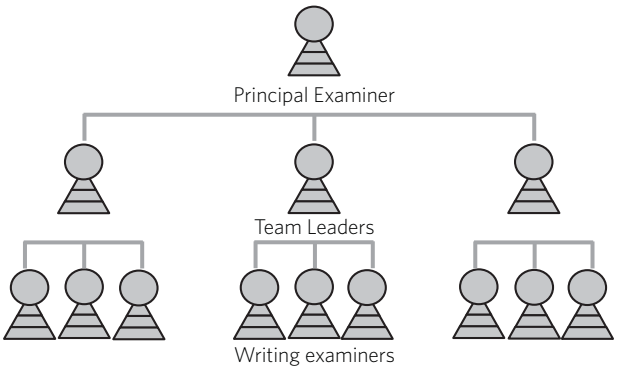


Once an examiner has been certificated and has registered availability they are then in a position to take on any marking opportunities that they are eligible for.

Performance feedback

In order to provide examiners with feedback on their performance, Cambridge ESOL makes use of its marking hierarchy (Figure 17) to monitor examiners during live marking, and the functionality of 'seeding'.

Figure 17: Examiner hierarchy



Monitoring

PEs and TLs have a key role in monitoring WEs during marking and the robust monitoring of the work of all examiners (WEs, TLs and PEs) is a crucial quality control measure, in line with Cambridge ESOL's view of best practice in examiner management. Monitoring and assessing the reliability of examiners is also a regulatory requirement (e.g. Ofqual, ALTE Code of Practice).

The purpose of Monitoring is twofold: a) to ensure candidates receive a mark which accurately reflects their ability; b) to provide WEs with ongoing support designed to help them maintain the assessment standards established during the training and certification stages.

All WEs are monitored by TLs, and TLs by PEs, throughout the marking period. The monitoring process covers the requirements for accuracy and consistency of marking and all aspects of administration.

TLs sample WEs' assessments of scripts at appropriate intervals during the marking period, checking that the mark scheme has been applied correctly. Any issues that arise are discussed between the TL and the WE and, if necessary, escalated to the PE.

In addition to Certification of Assessment, monitoring the work of WEs allows Cambridge ESOL to refine writing panels so that only the most competent examiners are used, and to identify WEs who could potentially become senior examiners on future panels.

Seeding

A process of seeding is available in scoris® Assessor, whereby a number of candidate scripts are given to all examiners to mark. This function can be used in a number of ways.

1. The marks awarded by each examiner can be compared with an agreed 'gold standard' mark in order to provide information on examiner marking variations. This helps to inform TLs and PEs during their monitoring of live marking,

i.e. rather than relying on their subjective judgement, they are able to compare WEs' marks on seeded scripts with the gold standard mark for those scripts.

2. Seeding makes use of live scripts from the session. This ensures that examiners are unaware of the fact that they are being used for monitoring purposes and therefore singled out by examiners for special attention.
3. *Post hoc*, information on each examiner's marking tendency at different intervals in the marking period can be determined, adding to the overall picture of an examiner's accuracy.

Feedback to examiners

Having engaged in marking, examiners are provided with feedback on their performance. This is different from the feedback that is given in the Certification of Assessment stage, where examiners are advised that they have completed the exercise and are therefore eligible to mark (or not). This is feedback given as a result of live marking. Having successfully completed the Certification of Assessment process, examiners at all levels are deemed to be competent, i.e. Satisfactory – so they are only contacted if there is an issue with some other aspect of their role. On the advice of the TL/PE, Cambridge ESOL follows up on issues such as WEs being unable to complete their marking target or slow rates of marking and, if necessary, an examiner will be withdrawn from marking.

Conclusion

This article has described the Quality Assurance procedure for Cambridge ESOL Writing examiners. It shows principally how throughout the process measures are taken to support examiners to ensure only the most reliable examiners are invited to mark. It also highlights the efficiencies that can be made from having training and certification delivered online.

References and further reading

Association of Language Testers in Europe (ALTE) (1994) *Code of Practice*, available online: www.alte.org/cop/index.php

Cambridge ESOL (2011) *Principles of Good Practice: Quality Management and Validation in Language Assessment*, available online: www.cambridgeesol.org/assets/pdf/general/pogp.pdf

Cambridge ESOL (2012) *Handbooks for Teachers*, available online: www.teachers.cambridgeesol.org/ts/exams/

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

ffrench, A (2003) The change process at the paper level, Paper 5, Speaking, in Weir, C and Milanovic, M (Eds) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002,* Studies in Language Testing volume 15, Cambridge: UCLES/ Cambridge University Press, 367–472.

Ofqual (2012) *Office of Qualifications and Examinations Regulation: Ofqual*, available online: www.ofqual.gov.uk

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.

Taylor, L and Galaczi, E (2011) Scoring 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, 171–233.

Perceptions of authenticity in academic writing test tasks

GRAHAM SEED ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ESOL

Introduction

This paper is based on a Master's dissertation submitted to Lancaster University, UK in December 2011. The dissertation was supervised by Dr Judit Kormos.

Authenticity is a frequently used concept when validating specific-purpose language tests: one such type is a test of writing for academic, or university admission, purposes. Cambridge ESOL claims to provide a range of such tests through its *IELTS*, *Cambridge English: Advanced (CAE)* and *Cambridge English: Proficiency (CPE)* tests, as well as previously through *IELTS*'s predecessor, the *ELTS*, which was specifically designed for academic purposes (Davies 2008:28–52). The tests *Cambridge English: Advanced* and *Cambridge English: Proficiency* were designed to test general English, as opposed to the English for Academic Purposes (EAP) *IELTS* test (Shaw and Weir 2007:12). However, *Cambridge English: Advanced* and *Cambridge English: Proficiency* are now, in addition, considered appropriate for 'Academic and Professional English' purposes (Cambridge ESOL 2010).

Authenticity in academic writing tests is often measured by discovering what the real-life target language use (TLU) situations in a university are for example by asking university lecturers for their perceptions of what the TLU situations are, and whether certain tests are appropriate. However, few studies have investigated perceptions from other stakeholders, notably English as a Foreign Language (EFL) students preparing to take such tests and EFL teachers of test

preparation courses. This study sought to obtain perceptions from these 'front-end' users about selected writing tasks from *IELTS*, *Cambridge English: Advanced*, *Cambridge English: Proficiency* and *ELTS*, and to compare these views with opinions of test developers at Cambridge ESOL. A task from *ELTS* was chosen for investigation as it exemplifies alternative features of academic writing tasks such as the subject-specific topic, integrated skills and note-taking elements.

Literature review

Bachman and Palmer (1996) include authenticity as one of six key elements to decide if a test has 'test usefulness'. Authenticity is also a 'central issue in validating direct assessment of writing' (Moore and Morton 1999:76). However, there are many differing perspectives on what authenticity comprises. Bachman (1990) distinguished between *situational authenticity*, the match of test tasks to real-life tasks, and *interactional authenticity,* the test taker's involvement and language ability in the test task. Read (1990:119), who was writing specifically about academic writing tests, included the notion of *predictive validity*: 'Are we [in a test] accurately predicting the extent to which individual students encounter difficulties in coping with the writing requirements of their academic courses?'

If academic writing tests are authentic, or at least *perceived* to be authentic, this increases the validity of the test. However, 'authenticity' and 'perceptions of authenticity' may be different things. The latter is a form of face validity, described by Ingram (1977:18) as the test's 'surface credibility or public acceptability'. Perceptions rely on people's opinions rather than facts, and may change frequently. Therefore, it is important for test developers to understand what current perceptions are, in order to make quality products, not only for test validity but also to meet market demand and expectations. It also encourages candidates to take the test seriously and perform to the best of their ability (Alderson, Clapham and Wall 1995:173). Positive perceptions of a test can therefore play an important role in a validity argument. Cambridge ESOL claims this to be true of their products too: see Saville 2003, and more recently:

> The organisation needs to be highly receptive to the needs, opinions and knowledge of our stakeholders. We aim to provide effective channels for two-way communication, listening to our customers and to other experts in language assessment and adjacent fields, and actively seeking the views of all our stakeholders (Cambridge ESOL 2011b).

As Lewkowicz (2000) explored, candidates may even be satisfied with taking *any* test, generally ignoring whether or not it can benefit them in the TLU situation. From her study, she stated that 'few saw aspects of authenticity or lack of it as important' and that test takers 'may view multiple-choice tests as authentic tests of language in contrast to tests of authentic language' (2000:59). Therefore, as long as a test is *perceived* to be academic by a test taker, any test may be considered authentic:

> Test-takers' perceptions of authenticity vary . . . Authenticity would appear not to be universally important for test-takers. On one hand these results are in line with Bachman and Palmer's (1996) notion that stakeholders' perceptions of test authenticity differ not only across but also between groups of stakeholders; at the same time they suggest that there may be a mismatch between the importance accorded to authenticity by language testing experts and other stakeholders in the testing process (Lewkowicz 2000:60).

It would be useful to see if these two points regarding stakeholder perceptions, summarised in Bachman and Palmer (1996:24), hold true more than 10 years after Lewkowicz's (2000) study, and for Cambridge ESOL products.

Research questions, methods and materials

The original dissertation (Seed 2011) firstly examined which general tasks and skills were needed for an academic Writing test task to be perceived as authentic, but this paper will only focus on the dissertation's second research question: *To what extent are various tasks in academic Writing tests provided by Cambridge ESOL perceived as authentic for the TLU?* As a result of this, a further question is whether there are differences in opinion between and within the three groups of test developers, students and teachers. This paper will focus on five selected tasks from tests which Cambridge ESOL has deemed appropriate to be used in making decisions for academic admission (see tasks in Appendix 1). Two tasks are from *IELTS* Academic Writing: summarising information from a bar chart in Task 1, relying on candidates processing visual information to guide their composition; and a Task 2 essay, containing very little input – a proposition and a rubric. Another task is from the revised *Cambridge English: Proficiency* for 2013: this is a compulsory essay where the candidate must read two short texts, and then summarise the main points while adding their own ideas into an essay. One task is from the current *Cambridge English: Advanced*, chosen deliberately as one, from the researcher's point of view, that is not overtly academic in focus, in order to test one of Lewkowicz's (2000) findings, as discussed above, that if candidates are told that any given test task may be used for academic purposes, they would accept it as academically authentic. It should of course be noted that *Cambridge English: Advanced* does contain other more academically focused tasks, and that this is just one task out of a whole test in which different constructs are being measured: not only academic but general and professional ones too. It should also be borne in mind that *Cambridge English: Advanced* is being revised for 2015, and that this particular task comes from the pre-revision version. A final task, which could be deemed to be the most academically authentic, is taken from an *ELTS* paper included in Davies (2008:192, 200). Here, the writing task is integrated with, or more precisely follows on from a reading activity. After reading a passage from a source booklet and answering multiple-choice reading comprehension questions relating to the text, the first writing questions are to draw a flow chart of the processes described in the passage and then tabulate the defects and causes of this process.

In this study, both focus groups and questionnaires were used to collect data, with the hope that the different collection methods would complement each other and become a

type of 'mixed-methods research', as discussed by Dörnyei (2007:163). Focus groups are often used in the first stages of a study, in order to use the resulting analysis to develop questions for a later questionnaire stage (O'Brien 1993). Therefore, in the final analysis, quantitative data obtained through questionnaires was used alongside the qualitative data from the focus groups.

Four focus groups were held during August 2011. Each comprised four or five participants and lasted approximately 45 minutes. Participants discussed the authenticity of the five tasks, and were informed that each task came from a test that could be used for university entrance purposes, but were not told exactly which test they came from. The first focus group was with Assessment Managers of *IELTS* and the *Cambridge English: Preliminary (PET); Cambridge English: First (FCE); Cambridge English: Advanced (CAE); Cambridge English: Proficiency (CPE)* tests from Cambridge ESOL. This group's members covered the term 'test developers'. The second and third groups comprised students studying in *IELTS* and the aforementioned ESOL examinations test preparation classes at a language school in Cambridge. The final group interview was with teachers of *IELTS* and the ESOL examinations test preparation classes at a language school in Cambridge. Discussions among the four focus groups were transcribed and coded according to categories which later helped form the questions in the questionnaire as described below (authenticity of topic, authenticity of skills and suitability for an academic Writing test of each of the tasks).

In the questionnaire phase which followed, respondents were asked to rate each task on a seven-point scale from zero to six in three criteria: *authenticity of topic,* defined as whether the topic was similar to a real writing task at university; *authenticity of skills*: whether the skills needed to complete the task were similar to those needed when doing real writing tasks at university, and *suitability for a test*: whether it was thought that the task was suitable as a university entrance test, regardless of whether the task was perceived to be authentic. The test developers' questionnaire was sent to internal staff and to external consultants of Cambridge ESOL. The students' questionnaire was sent to a variety of English language learners who were known to have taken or were due to take an *IELTS* or *Cambridge English: First, Cambridge English: Advanced* and *Cambridge English: Proficiency* examination.

They were invited to rate the authenticity of each task on a scale from 0 to 6, where 6 is extremely authentic, and 0 not authentic at all. Questionnaire respondents were also invited to make optional comments on the tasks, and this qualitative data was also coded. Descriptive statistics were run on the quantitative data using SPSS. T-tests were also run to compare the means between the two groups by seeing whether any difference in means was significant (p = <0.05). The results are shown in Figures 1–5.

Discussion

The findings of the quantitative data from the questionnaires are presented for each task with a summary of the major ideas from the focus groups when discussing each task, as well as any salient points from qualitative data from the questionnaires. The tasks are shown in Appendix 1. Two overriding points should be remembered: firstly, students gave far fewer qualitative responses, both in the focus groups and in the questionnaires. This may be due to the fact that the multilingual group had to express their feelings in English rather than their native language, despite all participants in the focus groups and questionnaires being approximately Upper Intermediate level or above. Secondly, it was noted that test developers and teachers thought about the expected language output of the tasks more than the students did, which is perhaps to be anticipated given the professional knowledge test developers and teachers have about language acquisition.

IELTS Academic Writing Task 1 – Summarising information from a bar chart

Teachers and test developers were impressed by the range of skills and language elicited, mentioning extrapolating information, making comparisons, showing change over time, organising ideas, summarising and also testing cognitive ability, all of which were seen to be authentic academic writing skills. Both groups also discussed the short word limit, which while inauthentic does encourage the authentic skill of extracting only the main ideas. Stronger candidates would be more successful at this, and so the task was therefore believed to be able to discriminate between test takers well. Teachers mentioned that a truly authentic task would actually go

Figure 1: Comparison of means for questionnaire responses: *IELTS* Academic Task 1 (Bar chart input)

IELTS Academic Task 1 (Bar chart input)		N	Mean (to 2dp)	Mode	Min	Max	SD (to 2dp)	T-test probability (to 3dp)	Difference in means for significant results (to 2dp)
Authenticity of Topic	TD	27	3.59	3	1	6	1.22	NS	
	SS	26	3.88	5	1	6	1.42		
Authenticity of Skills	TD	27	4.37	4	2	6	0.93	NS	
	SS	26	4.31	5	2	6	1.26		
Suitability for a Test	TD	27	4.41	4, 5	2	6	1.05	NS	
	SS	26	3.92	4, 5	1	6	1.41		

TD = Test Developers; SS = Students; NS = not significant; dp = decimal places.

Difference in means for significant results: + means test developers perceived this as more authentic; – means students perceived this as more authentic.

Figure 2: Comparison of means for questionnaire responses: *IELTS* Academic Task 2 (Essay)

IELTS Academic Task 2 (Essay)		N	Mean (to 2dp)	Mode	Min	Max	SD (to 2dp)	T-test probability (to 3dp)	Difference in means for significant results (to 2dp)
Authenticity of Topic	TD	27	2.37	1	0	5	1.50	$p = 0.000$	−1.70
	SS	27	4.07	5	2	6	1.38		
Authenticity of Skills	TD	27	4.00	3, 5	2	6	1.11	NS	
	SS	27	4.11	5	2	6	1.31		
Suitability for a Test	TD	26	3.77	5	0	6	1.37	NS	
	SS	27	3.96	5	0	6	1.76		

further by asking candidates to critique where the information came from, which would produce different language. Furthermore, both test developers and students believed that the task can be applied to students of all subjects, and so would be suitable for an academic Writing test that covered all disciplines.

Conversely, the quantitative data from students was negative, ranking this task last out of all tasks for topic and for suitability (see Figure 1). A possible explanation for this might be that more of the focus group students were studying for *IELTS* and were therefore accustomed to this type of task, whereas the questionnaire respondents had more experience with General English exams.

IELTS Academic Writing Task 2 – Essay

Test developers, students and teachers all saw that the basic skills used here of giving opinions, acknowledging the existence of other arguments, and weighing up the arguments to form a conclusion, are authentic for academic situations.

However, all commented that the topic was not particularly authentic, with students saying that more technical, specific questions would be more appropriate at university. It should be remembered, however, that *IELTS* is a test taken *before* university and therefore true authenticity is perhaps not necessary. Test developers and teachers commented that some candidates may lack knowledge and experience to answer this question, and that support in academic essays is usually found from external sources, but understood the demands of practicality in test situations, as the teachers discussed:

> 'You're still creating your own evidence, aren't you? You are creating your own evidence rather than drawing from a text or something concrete.'
>
> 'Yes, but I don't see how else you could practically do something like this.'

In this case authenticity has to be compromised in part by practicality because a timed test situation cannot replicate a real-life task of being able to refer to numerous sources over a much longer period of time before writing an essay.

The quantitative data supports these sentiments in that the mean for skills was much higher than for topic (See Figure 2). However, the task elicited large standard deviations, suggesting an amount of disagreement within the groups.

Revised Cambridge English: Proficiency Writing Part 1 – Summarising essay

The skills of synthesising, evaluating, summarising and adding to the information, all present in this task, were seen as very authentic by all respondents, as one student positively commented:

> 'In academic research this is exactly what you have to do. You have, well here there is not a title . . . but you usually have a title, an issue, a problem, then you have to start thinking about it, then you have to read about what the others have said on this point. So you read articles, journals, books, and then you have to find your, to have an opinion, to write it, quoting what the others said. So you have to summarise others' opinion, and then put your own ideas . . . So I think taking this as an example; not exactly this one, but, with very long passages to read, that is exactly what the academic research you do.'

It became apparent during the focus groups that the skill of summarising was seen as highly authentic for the academic context, and that this task therefore, as it specifically includes the need to summarise, was viewed as very appropriate.

Both the students and the teachers made comments based on their own experiences studying at university that the lack of an essay title and the time pressure were inauthentic. Again, the practicality of a timed test situation means compromises may need to be made. However, this task was received very favourably by the questionnaire respondents

Figure 3: Comparison of means for questionnaire responses: Revised *Cambridge English: Proficiency* Task 1 (Summarising)

Revised CPE Task 1 (Summarising essay)		N	Mean (to 2dp)	Mode	Min	Max	SD (to 2dp)	T-test probability (to 3dp)	Difference in means for significant results (to 2dp)
Authenticity of Topic	TD	27	3.78	4	2	6	1.15	$p = 0.015$	−0.84
	SS	26	4.62	5	1	6	1.27		
Authenticity of Skills	TD	27	4.85	5	2	6	0.95	NS	
	SS	26	4.81	5	3	6	0.94		
Suitability for a Test	TD	27	4.44	5	1	6	1.25	NS	
	SS	26	4.85	5	2	6	1.08		

Figure 4: Comparison of means for questionnaire responses: *Cambridge English: Advanced* Task 1 (Informal letter)

CAE Task 1 (Informal letter)		N	Mean (to 2dp)	Mode	Min	Max	SD (to 2dp)	T-test probability (to 3dp)	Difference in means for significant results (to 2dp)
Authenticity of Topic	TD	27	1.41	1	0	4	1.01	$p = 0.000$	−2.86
	SS	26	4.27	6	1	6	1.56		
Authenticity of Skills	TD	26	2.08	3	1	3	0.89	$p = 0.000$	−2.58
	SS	26	4.65	5	2	6	1.16		
Suitability for a Test	TD	27	1.78	1	0	4	1.09	$p = 0.000$	−2.57
	SS	26	4.35	5	1	6	1.41		

too (see Figure 3), backing up the feelings expressed in the focus groups, with both test developers and students ranking it highest overall.

Cambridge English: Advanced Writing Part 1 – Informal letter

Participants from all groups viewed this task negatively in terms of its academic authenticity, mainly seeing it as suitable for general purposes. While one of the student focus groups said they thought there were situations when they would write an informal letter at university, none of the participants who had already been to university could recall a time when they had actually done this. This disproves a theory that students would accept as authentic any task, even one that was not designed with academic purposes in mind, in a test that may be used for university entrance purposes.

On the other hand, participants in all groups found some academically authentic skills such as synthesising, organising, evaluating and recommending. Questionnaire responses from students were particularly positive, as seen in Figure 4. Perhaps this is due to *perceived authenticity* as described above, or perhaps because the questionnaire respondents had more *Cambridge English: Preliminary (PET); Cambridge English: First (FCE); Cambridge English: Advanced (CAE); Cambridge English: Proficiency (CPE)* experience than *IELTS* experience, making them more favourably inclined towards this particular *Cambridge English: Advanced* task. But on balance, as the teachers' focus group concluded, although it exhibits some academic writing skills, an informal letter writing task would not be wholly suitable for a test for academic purposes.

ELTS – Flow chart and table, after reading text

The test developers' group mainly discussed the integrated skills element of this task, claiming it to be authentic, but more a test of reading than of writing. This latter point was picked up on by teachers, saying that note-taking writing tasks are not authentic for tests as in real life this could be done in the student's L1. Test developers also wondered whether answering the reading questions correctly or incorrectly would have an effect on being able to answer the writing questions. One student focus group praised the task for its authentic summary-eliciting skills, while the other student focus group criticised its perceived difficulty, subject-specificity and the added complication of being an integrated skills task. As one student commented, although the task was authentic, it was not suitable for a test:

> 'I think this is more for the university, and not for the IELTS. Yes, while at university, not before.'

The messages from the focus groups have demonstrated that all three groupings were able to pick up on the reasons as to why *ELTS* developed into *IELTS* without integrated tasks, just as Davies (2008:64) commented that writing should not be linked to reading, because 'a weak writing performance therefore might be caused by a failure to understand fully the reading texts'. This perhaps shows that students and teachers, not just test developers, have an understanding of the challenges of authentic test design and delivery.

The quantitative data gives a varied picture, especially as the standard deviations are comparatively large for all three categories (see Figure 5). As opinion differed between and within the focus groups for each stakeholder grouping, so did they seemingly in the questionnaires too.

Conclusions

When comparing the test developer, student and teacher perceptions, the tasks broadly fell into three categories:

1. The *IELTS* and *Cambridge English: Proficiency* tasks: Generally speaking, there were no significant differences of opinion between the three groupings. All mainly agreed that these were relatively authentic academic tests. Useful

Figure 5: Comparison of means for questionnaire responses: *ELTS* (Integrated task)

ELTS (Flow chart and table after reading text)		N	Mean (to 2dp)	Mode	Min	Max	SD (to 2dp)	T-test Probability (to 3dp)	Difference in means for significant results (to 2dp)
Authenticity of Topic	TD	27	3.93	4	1	6	1.04	NS	
	SS	26	4.27	6	0	6	1.69		
Authenticity of Skills	TD	26	4.31	5	1	6	1.26	NS	
	SS	26	4.31	5	1	6	1.49		
Suitability for a Test	TD	26	3.31	3	1	5	1.19	NS	
	SS	26	4.00	5	1	6	1.65		

academic skills could be identified within these tasks. The only exception is that students were more satisfied with the authenticity of topic in each case, as shown by both qualitative and quantitative data.

2. The *Cambridge English: Advanced* task: In focus group interviews, there was general agreement that this specific informal letter task would be less suitable for academic purposes. Nevertheless, the students were much keener to search for and point out the positive features of why this task might be suitable and authentic for academic purposes, for example the skills it elicits. Questionnaire responses showed significant differences of opinion between students and test developers.
3. The *ELTS* task: There were no significant differences of opinion between the groups, but differences lay *within* the groups. Perceptions of authenticity for this task seemingly depended on personal preferences rather than according to stakeholder groupings.

Several noteworthy themes emerged as ones important to the debate between and within stakeholders about the perceived authenticity of academic Writing test tasks. To briefly summarise, they were:

- *General or subject-specific test tasks.* This is a major contention within specific-purpose language tests with regard to their content or topic. It was observed that subject-specific tasks were perceived to be more authentic, whereas tasks with more general topics were more practical, in that they were more accessible for all candidates, and it is the latter that seems to have had more influence, as evidenced by the revision of *ELTS* into *IELTS*. Many respondents, even the students, concluded that suitability overrides authenticity, especially because the test is taken *before* attending university rather than *at* university, meaning true authenticity would be inappropriate.
- *Finding the input or drawing on your own experience.* A candidate must call on content resources to complete the writing task, which either comes from input given or referred to in the task (*IELTS* Part 1, *ELTS*), or from the candidate's own prior experience and background knowledge (*IELTS* Part 2). Authenticity may favour the former, but practicality favours the latter. A solution, seemingly favoured by most participants in this study, is to merge the two together in the correct proportions as positively commented on in the revised *Cambridge English: Proficiency* Part 1, which combines an amount of input text and the instruction to include the test taker's own ideas.
- *Time and word constraints.* A notable and unexpected finding of this study was that the students attached significantly more importance to the need for time constraints as a reflection of authenticity, not just practicality, than teachers or test developers.
- *Integrated skills.* While the integrated natures of the *Cambridge English: Proficiency* and *ELTS* tasks were positively regarded as authentic, respondents did raise questions about the suitability and practicality of integrating skills in tests.
- *Summarising skills and tasks.* A frequently occurring theme was that tasks which involved summarising were always highly regarded as authentic by all stakeholders, but students in particular kept identifying this feature as frequently occurring in the TLU.
- *Authenticity or practicality.* One major conclusion that may be drawn from this research is that when considering authenticity, test developers, teachers and even students all naturally brought in practical considerations to the discussion, which by necessity may limit authenticity. Writing tasks may be academically authentic, but practicality of test administration imposes constraints. Authenticity is desirable but practicality must take precedence.

Lewkowicz's (2000) study suggested that students were not so interested in authenticity and were prepared to accept any test as suitable. Although this study had a slightly different focus, agreement can be made in part as students were generally more positive about all the tasks than the test developers were. Students were also more willing to find the positive and authentic points in the tests, while test developers and teachers were more willing to find criticisms and faults. Perhaps test developers constantly have in mind an ideal test which they strive towards and their perspectives on test tasks are a result of this.

On the other hand, students did demonstrate an awareness of what constituted authenticity in academic writing situations, contrary to Lewkowicz's assertions. Perhaps students have gained a certain amount of *assessment literacy*, or knowledge of assessment principles, as Taylor (2009) discusses. This may be due to the fact that student participants have been exposed to such matters through their preparation and study for taking language tests.

Ideally, a balance needs to be found where a test task is as authentic and predictive as possible, while also being suitable for a wide target candidature with different needs. The marks derived from the candidate's responses can therefore be used to make generalisations about a candidate's ability to cope in the TLU environment. Cambridge ESOL is in a position to promote the revised *Cambridge English: Proficiency* Task 1 and other similar tasks as being the most able to strike the balances mentioned above, especially those of authenticity and practicality.

References

Alderson, J C, Clapham, C and Wall, D (1995) *Language Test Construction and Evaluation,* Cambridge: Cambridge University Press.

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

Bachman, L F and Palmer, A S (1996) *Language Testing in Practice,* Oxford: Oxford University Press.

Cambridge ESOL (2006) *Cambridge IELTS practice tests 5: Student's Book with Answers,* Cambridge: Cambridge University Press.

Cambridge ESOL (2008) *Certificate in Advanced English: Handbook for Teachers,* Cambridge: UCLES.

Cambridge ESOL (2010) *Cambridge English: Selecting International Students for Higher Education,* Cambridge: UCLES.

Cambridge ESOL (2011a) *Cambridge English Proficiency: Specifications and Sample Papers for Examinations from March 2011,* Cambridge: UCLES.

Cambridge ESOL (2011b) *Principles of Good Practice: Quality Management and Validation in Language Assessment*, available online: www.cambridgeesol.org/assets/pdf/general/pogp.pdf

Davies, A (2008) *Assessing Academic English: Testing English Proficiency 1950–1989 – the IELTS Solution,* Studies in Language Testing volume 23, Cambridge: UCLES/Cambridge University Press.

Dörnyei, Z (2007) *Research Methods in Applied Linguistics,* Oxford: Oxford University Press.

IELTS (2011) *IELTS test takers – academic writing sample,* available online www.ielts.org/test_takers_information/test_sample/academic_writing_sample.aspx

Ingram, E (1977) Basic concepts in testing, in Allen, J P B and Davies, A (Eds), *Testing and Experimental Methods,* Oxford: Oxford University Press, 11–37.

Lewkowicz, J A (2000) Authenticity in language testing: some outstanding questions, *Language Testing,* 17 (1), 43–64.

Moore, T and Morton, J (1999) Authenticity in the IELTS academic module writing test: a comparative study of task 2 items and university assignments, *IELTS Research Reports* 2, 74–116.

O'Brien, K (1993) Improving survey questionnaires through focus groups, in Morgan, D L (Ed.), *Successful Focus Groups: Advancing the State of the Art,* Newbury Park, CA, USA: SAGE Publications, 105–117.

Read, J (1990) Providing relevant content in an EAP writing test, *English for Specific Purposes* 9, 109–121.

Saville, N (2003) The process of test development and revision within UCLES EFL, in Weir, C and Milanovic, M (Eds) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examinations 1913–2002,* Studies in Language Testing volume 15, Cambridge: UCLES/Cambridge University Press, 57–120.

Seed, G (2011) *Perceptions of Authenticity in Academic Writing Test Tasks,* unpublished MA dissertation, Lancaster University.

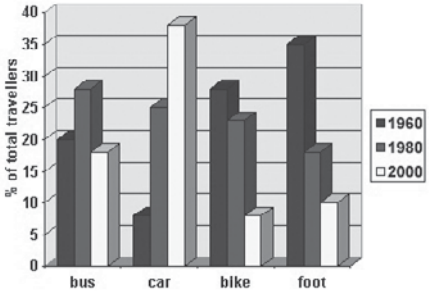
Shaw, S D and Weir, C J (2007) *Examining Writing: Research and Practice in Assessing Second Language Writing,* Studies in Language Testing volume 26, Cambridge: UCLES/Cambridge University Press.

Taylor, L (2009) Developing assessment literacy, *Annual Review of Applied Linguistics* 29, 21–36.

Appendix

***IELTS* Academic Writing Task 1 – Bar chart task (IELTS 2011)**

The graph below shows the different modes of transport used to travel to and from work in one European city in 1960, 1980 and 2000.



Summarise the information by selecting and reporting the main features, and make comparisons where relevant.

Write at least 150 words.

***IELTS* Academic Writing Task 2 (University of Cambridge ESOL Examinations 2006)**

Universities should accept equal numbers of male and female students in every subject.

To what extent do you agree or disagree?

Give reasons for your answer and include any relevant examples from your own knowledge or experience.

Write at least 250 words.

***Cambridge English: Proficiency* Writing (2013 Revision) Part 1 (Cambridge ESOL 2011a)**

Read the two texts below.

Write an essay summarising and evaluating the key points from both texts. Use your own words throughout as far as possible, and include your own ideas in your answers.

Write your answer in 240-280 words.

The Excitement of Advertising

Outdoor advertising has to attract, engage and persuade potential customers; it is the most important way of grabbing customers' attention and outdoor media continue to undergo a transformation. At the core of this transformation is the digital screen media, which encompass everything from giant screens to digital billboards. The technology is cheap and advertising agencies rave about the creative possibilities for advertisements which entertain, amuse, inform, make the environment brighter and enliven the world we live in.

Advertising: an undesirable business

Once upon a time outdoor advertising was straightforward. Posters were stuck up on anything from a bus shelter to a motorway hoarding. Many people considered this kind of advertising to be fairly dull, a harmless blot on the landscape and chose to ignore it. These people now regard digital advertising as a form of unwanted, creeping commercialisation: it attracts a buzz simply because it is new. They feel that any advertising which targets children or vulnerable adults is a dubious practice at the best of times, and digital advertising is, moreover, wasteful, damaging to the environment and completely unnecessary.

***Cambridge English: Advanced* Writing Part 1 (Cambridge ESOL 2008)**

Answer this question. Write your answer in 180-220 words in an appropriate style.

Last summer you had a job with an international company that organises music festivals. Your friend Jan has written to you asking about it.

Read the extract from your friend's letter and from your diary below. Then, using the information appropriately, write a letter to your friend saying whether or not you would recommend the job to your friend and giving your reasons.

Do you think I'd like the job? Most of all I want to hear plenty of music. I'd like to make enough money for a holiday too. If I could use my English and get useful work experience, that would be great!
Cheers,

Jan

July 2	Boring office work! No chance to learn anything. I answer the phone and make coffee.
July 10	Pay day! Things are improving! The money's not bad.
July 15	Did some translation and dealt with enquiries from English visitors.
July 22	Another free visit to festival!

Write your **letter**. You do not need to include postal addresses. You should use your own words as far as possible.

Write your letter. You do not need to include postal addresses. You should use your own words as far as possible.

ELTS Writing Question 1 (adapted from Davies 2008:192–200)

In this task you will first read a text about an engineering process known as 'Sand Casting'. Then you will answer some multiple-choice reading questions about the text.

[For the purposes of the focus group, the reading text and questions are not reproduced to their original size.]

Section 2: CASTING

The casting of liquid metal into a shaped mould and allowing it to solidify is a very convenient way of making solid metal components. One of the oldest casting techniques is *sand casting*. A mould is made by ramming moulding sand (basically, a silica sand with a proportion of clay as a binding agent) around a pattern of the part to be made. The pattern, which is generally made of hard wood, has to be made somewhat larger than the required dimensions of the finished casting, in order to allow for contraction of the casting during cooling. The mould is made in two or more parts, in order to facilitate removal of the pattern, and feeder channels, gates, and risers must also be incorporated in the mould. Hollow castings may be made by fitting cores in the mould. Cores, which have to be strong enough to be handled, and also to be able to remain largely unsupported within the mould, are often made from sand bonded with linseed oil, or made by the shell moulding process from sand-resin mixes. When the completed mould (and cores, if applicable) has been assembled, it is ready to receive the liquid metal. Liquid metal is carefully poured into the mould and allowed to solidify. When the metal has completely solidified the sand mould is broken up and the casting removed. *Fettling*, the operation to remove feeder heads, runners, and riser heads, is then carried out, followed by any necessary machining operations and inspection.

Owing to the low thermal conductivities of moulding sands, the rate of solidification within a sand mould is fairly low, and this results in casting possessing a fairly coarse crystal grain structure. Most metals undergo and considerable volume shrinkage during solidification, and it is the function of the riser heads to provide reservoirs of liquid metal to feed this shrinkage. Adequate provision of risers should largely eliminate the possibility of major solidification shrinkages zones within the casting, but finely divided inter-dendritic porosity is inevitable. Other defects which may occur in sand casting are sand inclusions, cold shuts, hot tears, and gas porosity. The major cause of sand inclusions within a casting is the washing away of loose sand from the walls of a poorly prepared mould. Cold shuts within a casting are a sign that the metal was poured at too low a temperature. Hot tearing, the fracture of a portion of the casting within the mould, is a result of tensile stresses being built up in parts of the casting due to thermal contraction, and is usually due to poor design of the casting. The causes of gas porosity within the casting may either be pouring liquid metal with a high dissolved gas content, or the generation of steam within the sand mould. This second type of porosity, known as reaction gas porosity, may occur when either the sand mould is too moist, or if the mould permeability is too low to allow any steam generated within the mould to escape to atmosphere.

Despite its apparent disadvantages, sand casting is suitable for the production of casting in almost any metal and of almost any size from a few grammes up to several hundred tonnes.

Read quickly through Section 2 and then answer these questions:

9. Why is a mould made in at least two parts?
A) to allow for expansion
B) to ease removal of the mould
C) to allow for contraction
D) to make removal of the pattern easier

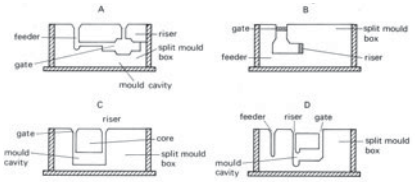
10. Why is linseed oil mixed with sand during some casting processes?
A) to strengthen a core
B) to support a mould
C) to make a casting stronger
D) to stop a core melting

11. The process known as 'fettling' takes place
A) when a metal solidifies.
B) before a mould is removed.
C) after a mould is dismantled.
D) as more liquid metal is added.

12. The 'coarse crystal grain structure' is due to the
A) high conductivity of the sand used.
B) construction of the sand mould.
C) pace at which the metal solidifies.
D) slow rate of the casting process.

13. 'Riser heads' are designed to
A) reduce the volume of liquid metal.
B) provide escape channels for excess liquid metal.
C) remove the impurities in the liquid metal.
D) increase the capacity for liquid metal.

14. Which of the diagrams best illustrates a simple sand casting mould?



Then, referring to this text, answer the following two writing questions

a) By means of a flow chart show the various stages in the sand casting process.

b) Tabulate the possible defects and their causes which may arise during the sand casting process.

A comparability study of computer-based and paper-based Writing tests

HEIDI ENDRES ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ESOL

Introduction

This paper is based on a Master's dissertation submitted to the University of Leicester in 2010. The dissertation was supervised by Glenn Fulcher.

As a result of developments in technology, assessment providers have found themselves faced with a demand for tests delivered on computers in addition to the traditional paper and pencil format. While the growth of computer-based testing (CBT) has led to exciting developments, such as computer-adaptive testing, and allowed for more ease and flexibility of delivery, it has also naturally raised many issues.

By far the most important issue is whether the mode of delivery affects the validity and reliability of the test. While in the future CBT will almost certainly become the more usual mode of delivery, computer-based (CB) and paper-based (PB) tests are likely to co-exist for a while, so equivalence between the two is very significant (McDonald 2002). The extent to which CB and PB testing measure the same trait determines whether they can replace each other (Neuman and Baydoun 1998:72).

Literature review

The comparability of PB and CB writing

While the differences for the test taker between writing on a computer and on paper (apart from the obvious effect of keyboard skills) may seem less problematic than the many issues raised when comparing the test-taking experience for a candidate taking PB and CB Reading or Listening tests, there are still many questions that need to be addressed. Do candidates make more errors in one mode or the other? Do they produce different language? Do they use different cognitive processes? How does rating across modes compare?

While there was found to be no difference between modes in the Cambridge ESOL trials for the CB IELTS Writing test (Green and Maycock 2004), some researchers have found that students may be disadvantaged when a CB option is not offered and they are required to take the PB version (Bocij and Greasley 1999, Russell and Haney 1997). Now that more and more students primarily write only on computers, it is possible when taking a PB writing test that the texts they produce are qualitatively different from those produced on computer because of lack of familiarity with this way of writing. Russell and Haney (1997: 2000) carried out two studies in US schools and in both cases, students performed better on computer. They claim that 'The situation is analogous to testing the accounting skills of modern accountants, but restricting them to the use of an abacus for calculations' and, consequently, they wonder whether writing on paper is now less of a real-world task (2000:2).

Others are asking similar questions (Lee 2004, Li 2006). Whereas in the past the question was whether it was fair to test candidates on computer, now people are asking if it is fair to force candidates to use a pen when they do most or all of their course writing on computer (Li 2006), and whether this affects the validity of the test. Lee (2004) suggests that as computers become the more usual medium for communication in writing, PB tests could become less authentic and potentially unfair. This is certainly an issue that needs to be considered.

What and how candidates write

One thing which stands out when looking at research on the texts candidates produce in both modes is their length. Several studies have found that students tend to write more on computer tests (Horkay, Bennett, Allen, Kaplan and Yan 2006, Li 2006, Russell and Plati 2002, Wolfe, Bolton, Feltovich and Niday 1996). Russell and Haney (1997), in their aforementioned study, compared student performance on CB and PB Writing tests and found that students wrote almost twice as much on computer.

Although candidates generally produce more text on computer, this does not necessarily mean that the quality of what they produce is better. According to some research, the linguistic features and complexity of the writing of school children appear to be the same in both modes (Peak 2005, Russell 1999, Russell and Haney 1997), although Wolfe et al (1996) found that computer experience had an effect on the texts produced, with less experienced computer-users writing shorter and simpler sentences on a CB test and more experienced users writing shorter and simpler sentences on a PB test. However, most research on PB and CB Writing tests concentrates on score comparability and less appears to have been done on the actual features of the texts produced. A notable exception to this is Chambers (2008), who studied the textual and linguistic features of CB and PB *Cambridge English: Preliminary (PET)* candidates' scripts. She noted that CB texts showed more lexical variation; that the CB texts had fewer paragraphs (38% of CB candidates wrote only one paragraph, as opposed to only 5% of the PB candidates); there were different punctuation issues in the two modes; and candidates produced different types of lexical errors. Chambers concludes that some of these differences may be attributable to the candidates' familiarity with email and chat and word processing tools.

The writing process

One of the great advantages of technology and CB testing is that it allows us to trace the writing process of candidates more easily and see exactly how they produce their texts

and what revisions they do as they are writing. Li (2006) compared the revisions of ESL students in PB and CB Writing tests and found that the CB students did significantly more revision. She also noted that the type of revision was different. There were many corrections of typing errors in the CB texts, but few corrections of 'slip of pen' errors in the PB texts. Students writing on computer also inserted sentences throughout their texts as they were writing and revising, whereas the students writing on paper only inserted sentences at the end of a paragraph.

In a study of the individual composing processes of six Korean students of English when writing on computer or paper, Lee (2002) found differences in planning, text production and revising across the two modes. She found that all these processes were more interwoven on computer than paper and there was significantly less pre-writing time, although there was no significant difference in scores. However, it should be noted that Lee's sample was very small.

The rating of PB and CB scripts

When it comes to the comparability of PB and CB writing tests, it is not simply a matter of how the candidate performs, but also of whether there is equivalence in the rating of the two modes. In fact, this appears to be quite an important issue as research suggests that candidates taking Writing tests on computer receive lower marks than those taking PB Writing tests (MacCann, Eastment and Pickering 2002, Shaw 2003). This seems surprising at first as you would expect poor handwriting to prejudice raters and potentially cause problems if a text is hard to read. However, it is generally thought that errors are more clearly seen in typed texts (Chambers 2008, MacCann et al 2002). Other possible reasons are that a handwritten text feels more personal to a rater, or that typed texts look shorter (MacCann et al 2002). Chambers (2008) suggests that the surface features like fewer paragraphs and misuse of capitalisation in CB scripts may also affect raters' perceptions of the texts.

It should be noted, however, that not all research shows a tendency to favour handwritten scripts. Lee (2004), in a study of an ESL placement test, found exactly the opposite and claims that about a third of the 480 students would have been placed at a higher level if they had written their essays on computer, purely due to raters marking PB scripts more harshly than CB scripts.

Methodology

In this study texts written by EFL students under exam conditions on paper and computer were analysed to see if there was a significant difference in performance in the two modes. A short questionnaire was also administered in order to determine students' familiarity with computers.

The research questions were the following:

- Are there any differences in performance on computer-based and paper-based Writing tests for 12–16 year old learners of English?
- What are the specific differences?
- What can the differences be attributed to?
- Can the differences have an effect on candidates' scores?

The participants were 28 Spanish students doing a preparation course for the *Cambridge English*: *Preliminary (PET)* exam. Students of one nationality, age group (secondary school children) and level were chosen to avoid the data being influenced by factors such as native language, age and English language ability, insofar as that was possible.

Two different composition prompts were selected – one for the PB test and one for the CB test. They were both from released *Cambridge English*: *Preliminary* exams and they were both from Part 3, which requires candidates to write a short story. Two similar tasks were chosen to avoid any issues with task types affecting the students' performance.

In order to gain a general understanding of the group's level of computer familiarity and frequency of use, the students were asked to complete a short questionnaire. This also permitted the sample to be divided into two groups (high-frequency and low-frequency), which allowed the possibility of investigating whether familiarity or frequency of use had any noticeable effect on the students' scores, errors and grammatical cohesion in the two modes.

The students were divided into two groups of 14 and the order of administration was varied so that one group took the CB test first and the other group took the PB test first. After completing the two tasks, the students were then asked to complete the questionnaire.

All the tasks were rated by two independent, trained and highly experienced Cambridge ESOL *Cambridge English*: *Preliminary* Writing examiners, who rated the texts as they normally would according to the *Cambridge English*: *Preliminary* (2011) General Mark Scheme, which was the mark scheme used at the time for Part 3 tasks (please note that this has now changed to an aligned mark scheme, see Lim's article in this issue). The markers gave each script a mark out of 15. An average of the two raters' marks was then used for the study. Rater agreement was also calculated, which at 76% indicated a relatively high level of agreement.

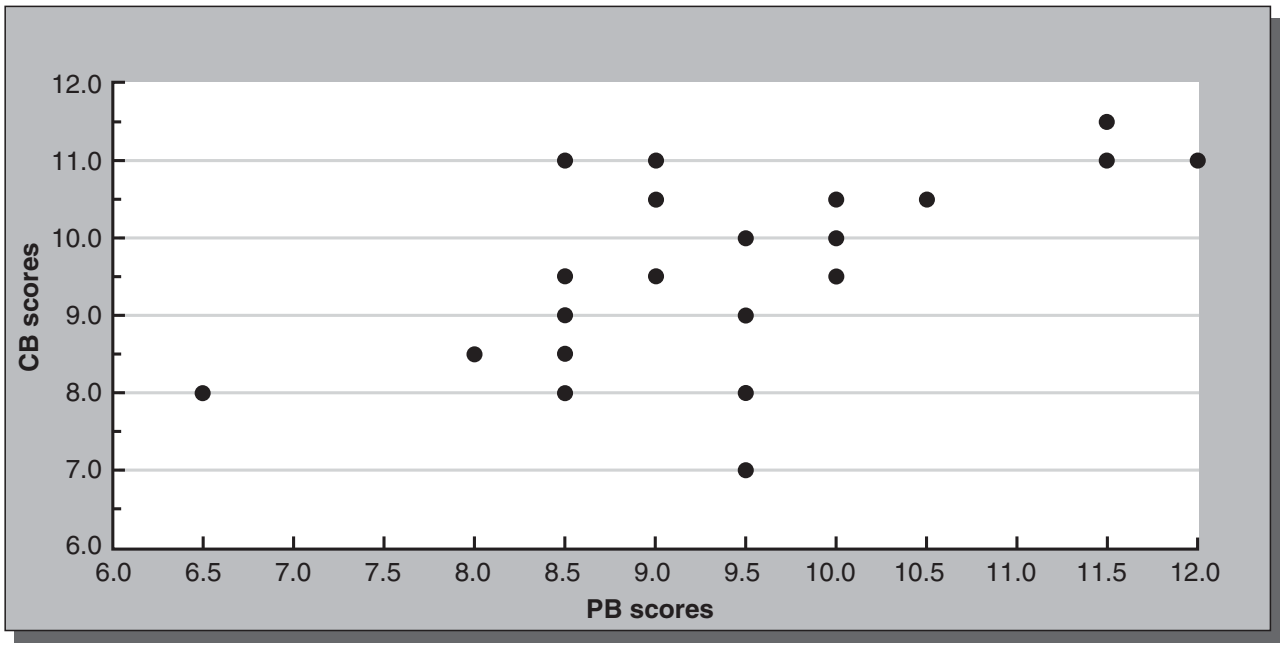
While the rater agreement is relatively high, Rater 2 uses a restricted range of the 15-point mark scheme, with nearly all the marks falling between 9 and 11. Rater 1, however uses a much broader range of 7 to 13.

The completed scripts were given to the markers without names and they were not told the nature of the study. They were not aware that they were marking two scripts from each student, one in each mode. It was hoped that this would minimise any unconscious bias on the part of the markers.

Results and discussion

Three analyses of the texts were then carried out. Firstly, a comparison of the marks and surface features (length of text, number of paragraphs, etc.). Secondly, an analysis of the errors. Thirdly, a discourse analysis of grammatical cohesion. As both the error and discourse analyses were subjective, a second rater was asked to code a sample of scripts and the two ratings were compared to estimate the coder reliability. The percentage coder agreement rate was then calculated

Figure 1: Scatterplot of scores under the two conditions



and found to be 74% for the error analysis, which indicates a relatively high level of agreement, and 98% for the discourse analysis, indicating a very high level of agreement. After comparing the errors and discourse markers in the PB and CB scripts, the findings were then related to the results from the questionnaire.

The initial analysis was a comparison of the students' marks in both modes and a comparison of surface features of the text, such as text length and organisation of the texts into paragraphs.

Scores and text length

Paired t-tests were used to analyse some of the data. This allowed comparisons between two means in relation to the data collected. For the first test, on scores out of 15, the mean was 0.34 higher for the CB test than the PB test (see Table 1 below). However, the standard deviation was almost identical.

The paired t-test results found a two-tailed p-value of 0.1214, suggesting there was no statistically significant difference in mean scores between the two conditions of computer and paper-and-pencil delivery.

Table 1: Mean scores under the two conditions

Test	PB	CB
Mean	9.339	9.679
SD	1.155	1.156
SEM	0.218	0.219
N	28	28

Table 2: Mean number of words under the two conditions

Test	PB	CB
Mean	110.11	137.46
SD	15.60	52.62
SEM	2.95	9.95
N	28	28

The scatterplot (Figure 1) shows the relation between the CB scores and the PB scores.

The correlation between the CB and PB scores was calculated, which at .58 (28% shared variance) shows only a moderate relationship. It would seem that candidates in the mid-ability range received inconsistent scores under the two conditions, as we can see from the scatterplot. This could be due to several reasons. This is a small and fairly homogeneous sample, as shown by the small standard deviation (see Table 1). Had the sample been larger and more varied, a higher correlation could have been expected. Moreover, there are two outliers, which reduced the correlation. One candidate scored 9.5 on the PB test and only 7 on the CB test, and another scored 6.5 on the PB test and 8 on the CB test. The other possible cause of the moderate correlation could be due to the fact that Rater 2 uses only a restricted range of the scale, as discussed above.

For the second t-test, on the number of words, the mean was much higher for the CB test than for the PB test: 27.35 more words were used on average. There was also a far greater difference between the standard deviations of the two tests. The standard deviation of the CB test was nearly 53, whereas for the PB test it was under 16 (see Table 2). Overall, on the CB scripts, students used more words, and there was much greater variability in word length.

The paired t-test results for this test found a two-tailed p-value of 0.005, which is statistically significant. Therefore the mode of writing made a difference to the length of the texts, with participants producing longer texts under the CB condition.

The difference in the length of the texts confirms previous findings (Horkay et al 2006, Li 2006, Russell and Haney 1997, Russell and Plati 2002, Wolfe et al 1996) and would suggest that students appear to take advantage of being able to use a computer by writing more.

Regarding the organisation of the texts, it was interesting to note that while 36% of students wrote two paragraphs on the paper-based test, none of them wrote more than one paragraph on the computer-based test, even though the texts were generally longer.

Error analysis

This analysis was based on the method described by Ellis (1994) with the errors being divided into morphological, syntactic, lexical and mechanical errors. Bardovi-Harlig and Bofman (1989) used the first three of these categories in their study of the errors of learners taking a university placement exam. In their study they discounted spelling errors, but it was felt this was particularly relevant to the research due to the potential for students to commit 'typos'. Therefore, the category of mechanical errors was added according to D Ferris's classification (2002). Punctuation errors were also included in the category of mechanical errors.

A third t-test was used to compare the number of errors in each group of scripts. The errors were counted and then the total number of errors for each student was calculated as a percentage of the total words in each script. For the t-test on errors, the mean was 0.35 higher for the CB test than the PB test (see Table 3). However, again the standard deviation was very similar.

The paired t-test results found a two-tailed p-value of 0.7263, suggesting there was no statistically significant difference and that writing a composition on computer or by hand did not have any significant effect on the number of errors produced.

Table 3: Mean number of errors under the two conditions

Group	PB	CB
Mean	13.7193	14.0714
SD	5.1212	5.2670
SEM	0.9678	0.9954
N	28	28

The next stage was to look at the type of errors the students committed in each mode, to find out if there was any difference in the type and/or distribution of errors in the CB and PB tests. The scripts were coded and the number of error types was counted. The results of the error analysis can be seen in Table 4.

Table 4: Analysis of errors under the two conditions

Error type	Total PB	% of total PB errors	Total CB	% of total CB errors
Morphological	220	52.38	257	47.94
Syntactic	45	10.71	37	6.9
Lexical	51	12.14	67	12.5
Mechanical	104	24.76	175	32.64
Total errors	420		536	
Total words	3,083		3,849	

As can be seen from this data, the distribution of errors is fairly similar. The only noteworthy difference is the number of mechanical errors. These were divided into spelling errors and punctuation errors. One prominent difference is errors of capitalisation. These errors accounted for 2.62% of the total PB errors, but 8.58% of the CB errors. Particularly noticeable was students' use of lower case 'i' for the first person singular pronoun when they were writing on computer. This was very common and there were very few instances of the same error in the handwritten samples, which is consistent with Chambers' findings (2008).

While the number of spelling errors committed in both modes is similar, the types of spelling errors committed are very different. In the scripts written on paper, the errors the students made are common spelling errors for learners, and particularly for Spanish learners of English. There are many examples of misspellings due to L1 graphology or phonology, for example, such as 'diferent', 'nervious' or 'sed' instead of 'said'. In the scripts written on computer, however, a large amount of the orthographic errors appear to be 'typos' caused by using the keyboard. For example, words with two letters transposed ('frist' instead of 'first', 'frineds' for 'friends'); extra letters added ('firefrightened' for 'frightened', 'befofre' instead of 'before'); incorrect letters in words due to their close proximity on the keyboard ('wiyh' instead of 'with', 'nodody' for 'nobody'); letters missing at the beginnings and ends of words ('nteresting', 'artificia'). The many examples of these types of errors in the CB scripts and very few in the PB scripts suggests they can be traced to keyboard use.

It seems clear that using a keyboard had some effect on the students' spelling and punctuation. This contradicts Wolfe et al's (1996) study of the performance of secondary school students on a PB and CB Writing test, where he found the number of mechanical errors to be the same on computer and paper. However, other research has suggested that poor keyboarding skills can have some impact on the written texts students produce (Russell and Haney 2000, Wolfe and Manalo 2004). It has been shown that errors in typed scripts are more visible to the reader; however, future research may be needed to determine whether the presence or absence of these keyboarding and capitalisation errors would impact examiner ratings.

Discourse analysis

The aim of the discourse analysis was to investigate any differences in the number or type of grammatical cohesive devices used in one mode or the other. The number of reference words and conjunctions were counted. The results can be seen in Table 5.

Table 5: Number of grammatical cohesive devices under the two conditions

Grammatical cohesive devices	PB	CB
Reference	659	870
Conjunction	190	238
Total words	3,083	3,849
% of total words (reference)	21.37%	22.6%
% of total words (conjunction)	6.16%	6.18%

Once again the findings were that the scripts written by computer and by hand were remarkably similar. In order to see if there was any difference in the type of reference words used, they were divided according to Hasan's classification (1985) (see Table 6).

Figure 2: Scatterplot of the relation between the CB scores and wordcounts

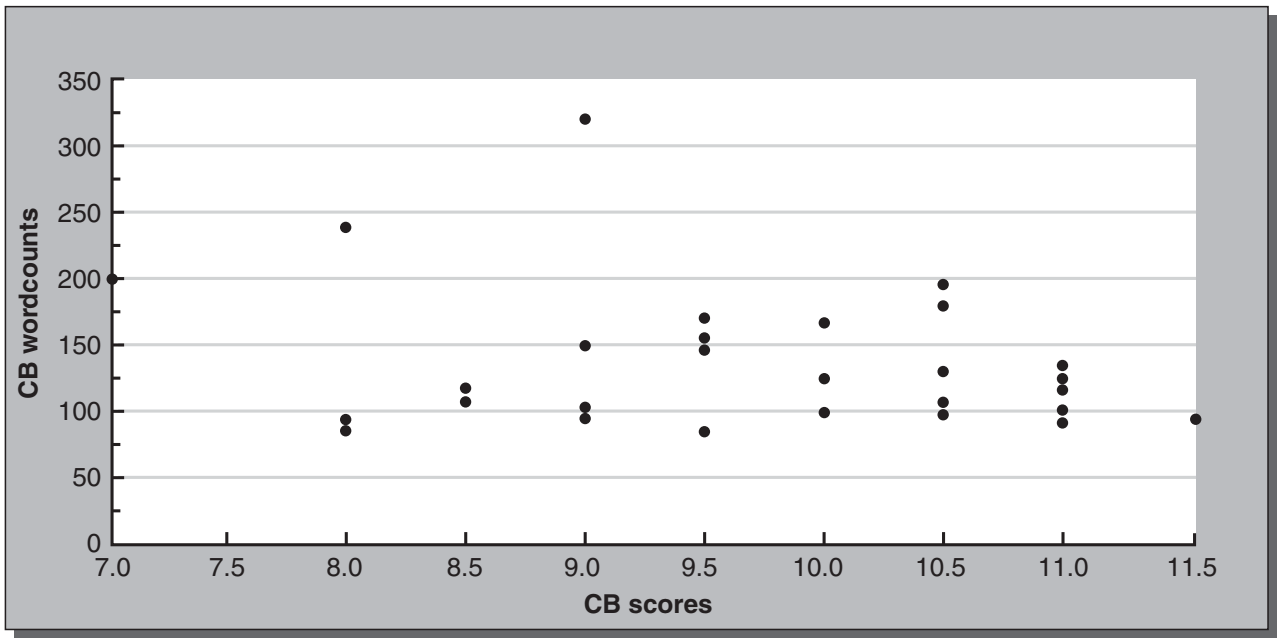


Table 6: Number of reference devices under the two conditions

Reference devices	PB	% of total PB words	CB	% of total CB words
Pronominal	513	16.63%	673	17.48%
Demonstrative	31	1%	28	0.02%
Definite article	100	3.24%	142	3.68%
Comparative	15	0.49%	19	0.49%

The results suggest that there is little difference in the actual structure of the texts produced by these students in either mode, although it should be noted that the discourse analysis was limited.

Questionnaire analysis

Finally the results were compared to student questionnaire responses. From that data, it was concluded that in general these particular students use a computer less often for formal writing, such as schoolwork and homework, and more often for leisure and informal communication. This means that, while they may be very familiar with computers and use them for several hours a day, they are perhaps less accustomed to using them for the type of writing required in an exam.

Candidates were also asked which mode they preferred, but while most preferred doing the exam on computer, it appears that their performance did not bear any relation to their preferences.

As preference seemed to be an unreliable predictor of score, the students were divided into two groups according to the frequency of their computer use in order to determine whether the students who used a computer more often performed better on the CB test than students who used computers less frequently. One group included the students who used a computer for more than five hours a week; the other included those who used a computer for less than five hours a week. The average PB and CB scores for each group were then calculated (see Table 7), and also the number of errors in each mode for each group.

Table 7: The effect of computer familiarity on scores and errors under the two conditions

	High frequency group	Low frequency group
Mean PB score	9.27	9.45
Mean CB score	9.63	9.75
PB errors (per 100 words)	13.2	14.66
CB errors (per 100 words)	14.06	13.2

There appears to be hardly any difference between the two groups. While the low frequency group did slightly better in both modes, the difference in average scores is minimal. Regarding the errors, it is actually the opposite to what would be expected, with the high frequency group making more errors on the CB test and the low frequency group making more errors on the PB test, although again the differences are too small to be significant.

Further investigation

The results show only one statistically significant difference and that is the number of words produced in the two modes. However, this difference does not seem to have had any impact on the overall mean scores. It could be due to the greater variation in the number of words on the CB tests. The fact that the length of the students' texts had no effect on their scores is contrary to previous research, which tends to show that increased text length leads to higher scores. According to Chodorow and Burstein (2004), longer texts in the TOEFL exam generally receive higher marks from raters and they claim that this could be because in a longer text there is more opportunity in which to fulfil the marking criteria (range of vocabulary, development of arguments, and so on). To investigate this further and try to determine why there is such a great variation in text length, the number of words produced by the high and low frequency groups on the CB tests was compared to see if computer familiarity had any effect on this difference (see Table 8).

Table 8: The effect of computer familiarity on the length of CB texts

	High frequency group	Low frequency group
Mean wordcount	130.11	150.07

As can be seen from the table, the low frequency group produced on average 20 words more per script than the high frequency group. This is contrary to Wolfe et al's (1996) study of the influence of student experience with word processors on the quality of essays, who found that students with less computer experience wrote fewer words on the CB essays.

It was also decided to determine whether language proficiency had any impact and the students were divided into two groups according to their score, comparing the number of words produced in the CB tests by the top half scorers and the bottom half scorers (see Table 9).

Table 9: The effect of language proficiency on the length of CB texts

	Top half scorers	Bottom half scorers
Mean wordcount	126.36	148.57

This calculation shows that the mean wordcount is 22 words more per script on average for the bottom half scorers than the top half scorers. The scatterplot (Figure 2) shows the relation between the scores and the wordcounts.

Lower ability candidates producing longer texts is an unusual finding. In this case it could be due to the greater variability and the high standard deviation on the CB test. The scatterplot also shows that there is greater variability of text length among students at lower levels. This confirms existing research which suggests that students with a lower level of English are more likely to be negatively affected by computer-based delivery than higher level students (Clariana and Wallace 2002, Wolfe and Manalo 2004).

Conclusions

For this small sample, the results of the analysis showed that there was little difference in performance on the computer-based and paper-based test at a group level. In fact, apart from the number of words, paragraphs and mechanical errors (both spelling and punctuation), the texts produced by these students were remarkably similar in terms of accuracy and grammatical cohesion, as were the scores achieved in both modes.

The errors of punctuation and the lack of separate paragraphs in the computer-based texts could be due to the fact that the nature of writing is changing as a result of increased computer use. Technology is changing the way we think about writing and how we do it (Chambers 2008). This group of students frequently uses computers for communication: for email and chatting. The way young people write when writing emails and particularly when chatting is very different from how people write (or used to write) on paper: writing is less formal; shortened versions of words are used; people use 'textspeak' not only in text messages, but also in emails; people are less concerned about capitalisation, punctuation, accents. All those formal customs are becoming less relevant when it comes to writing on a computer (S Ferris 2002). This may also affect raters' perceptions of computer-based scripts – perhaps the mechanical errors are not penalised as harshly by the raters as they are only to be expected when writing on a computer.

While Goldberg, Russell and Cook's (2003) meta-analysis of studies on the effect of computers on students' writing found a significant mean in favour of quantity and quality of writing on computer, with this sample there appears to be very little difference in the actual quality of the texts they produced. Apart from the mechanical errors, there was little difference in the number or type of errors committed. The discourse analysis also showed a strong comparability between modes, suggesting that the actual texture of the texts is also very similar.

Regarding computer familiarity and use, it seems that the amount of time the students spend using a computer had little effect on their performance. Also it seems that preference for either mode bore no relation to the results the students achieved in the tests. From the questionnaire, the conclusion can be drawn that this particular group of students is probably equally familiar with writing on a computer and on paper, still mainly using pen and paper at school. Perhaps this accounts for their remarkably similar performance in both modes. Whereas in the past students produced better texts on paper than computer, now the trend in some places is that students produce better texts on computer. This particular group does not appear to be advantaged or disadvantaged by either mode.

However, as pointed out earlier, while the sample performed similarly as a group, there were individual differences among the students. Only four students achieved the same mark in both modes. Consequently 24 students received a different mark depending on whether they took the test on computer or paper. The differences may not necessarily have been due to the mode of delivery: perhaps they found one prompt easier than the other or suffered from fatigue while doing the second test. The fact is, though, that 18 students achieved a higher mark on the CB test and this may have been due to the mode. It is important when doing these studies not to rely solely on group results, but also take into consideration individual performance or variation. As Clariana and Wallace said 'though the difference may be small, the consequences for an individual student may be substantial (i.e. pass versus fail)' (2002:594). These findings, however, have found that this impact on the individual is not a significant factor in PB and CB testing of writing.

While my research has many limitations, and it may not be possible to extrapolate the results to other contexts, some findings are of interest. Firstly, although the mode of delivery showed no impact on scores, there were some differences in performance on the computer-based and paper-based test, which means that for whatever reasons, writing in the two modes is a different experience for candidates and should not simply be considered comparable. Secondly, these results (and all the conflicting results in this field) highlight the need for more research on the equivalence of PB and CB Writing tests.

The future

One of the main reasons why comparability studies should continue to be carried out is that the student population is changing and with technology becoming more widespread,

differences in performance will be continually changing as well. These results do not suggest that paper-based and computer-based writing are comparable, rather that they were comparable for this particular group, at this particular moment in time.

The impact of computer familiarity and anxiety on test performance is rapidly losing relevance. In fact, nowadays, as test takers are becoming more and more comfortable with technology, we are facing the opposite problem: that of candidates being disadvantaged by or anxious about taking a PB test (Chapelle and Douglas 2006). With this changing situation, it is essential that research is ongoing.

As technology becomes even more widespread, we will be faced with a whole new set of problems. When computers become a central part of schooling and the way students learn, assessing them in a different medium will not lead to valid and reliable judgements (Bennett 2002). There is now a large number of virtual universities and online courses, which bring education to those who may not otherwise have access – again, if students are doing a whole course on computer, how can we realistically expect them to be assessed on paper? Already, some previously mentioned studies (Russell and Haney 1997, 2000) have found that students' writing performance is being unfairly judged by PB tests as the use of computers in schools becomes more common. The method of assessment needs to reflect the tools employed in teaching and learning (Bennett 2002). At the moment, with the use of computers in schools varying so greatly, one possible way to eliminate unfairness would be to offer candidates a choice of which mode they take their exam in. However, Russell and Haney (2000) concluded that given the choice, students may make bad decisions, a fact confirmed by this study, which proves that preference is not necessarily an indicator of improved performance.

This leads to another point about authenticity. With students now increasingly using computers for writing, the authenticity of a test delivered on paper is threatened. And if students are used to writing on computers with all the functionality that normally entails, is it fair to deny them that functionality in assessments? Tools like cutting and pasting, block deleting, formatting, spell-checks and auto-correct are becoming a normal part of writing and have been shown to facilitate revision, editing and text generating (Li 2006). Should we start to allow these tools in assessment to more realistically reflect how students write nowadays? Naturally this raises many issues with a test of language, where correct spelling is tested, but actually, with spell-checks and the like, how important will it be for people to be able to spell correctly in the future? By depriving candidates of certain functionalities, are we denying them the opportunity to perform to the best of their ability?

There can be no doubt that computer-based testing is here to stay and the question is no longer whether to use computer technology in assessment, but how to integrate it more effectively (Liu, Moore, Graham and Lee 2002). However, there have been suggestions that PB and CB comparability studies, while advisable from the perspective of fairness, may limit innovation. This is particularly contentious when it comes to computer-adaptive tests, as they are thought to be superior to PB tests both in their administrative features and their psychometric properties and making them comparable restricts them and diminishes the advantages they bring (Wang and Kolen 2001). By ensuring that CB tests are comparable to PB tests, we are not taking advantage of all the possibilities that new technology has to offer and all the potential improvements to testing – the possibility of measuring skills and constructs that are beyond the capabilities of traditional paper-based assessment (Bennett 2002).

Bennett (1998) claims there will be three generations of CB testing. The first is when CB and PB tests will be very similar and CB testing will make limited use of technology, as is the case now. The second generation will be when we start to see new item formats, more use of multi media, perhaps automatic item generation. In the third generation he suggests that the distinction between learning and assessment will become blurred and simulation will be used. This idea serves to highlight how testing will change in the future. While at the moment PB tests are the benchmark for CB tests, in the future it is likely to be the other way around (McDonald 2002). And with new generations of CB tests, comparability studies will need to be done between one CB test and another CB test, to ensure that the new technological advances are measuring the same construct as before and not introducing construct-irrelevant variance into assessment. It will be a never-ending process of comparability and validation studies to keep up with technological changes.

But assessment needs to keep up with these changes and needs to reflect what and how students are learning, and the new skills they are developing. How relevant is it now to test how students process information from a printed text, when in real life they process most of their information from TV, radio and the internet (Bennett 1998)? Should we not be testing how well they use these real-life everyday skills? Not only are continuing studies in equivalence needed, but also research into new types of testing and new and innovative item types, in order to make the most of what computers have to offer. There can be no doubt, after all, that the future of assessment will lie in computer-based testing and we need to be ready when the time comes to throw off the shackles of the old paper-based tests.

References and further reading

Bardovi-Harlig, K and Bofman, T (1989) Attainment of syntactic and morphological accuracy by advanced language learners, *Studies in Second Language Acquisition* 11 (1), 17–34.

Bennett, R E (1998) Reinventing assessment: speculations on the future of large-scale educational testing, *ETS Research Reports*, available online: www.ets.org/research/policy_research_reports/pic-reinvent

Bennett, R E (2002) Inexorable and inevitable: the continuing story of technology and assessment, *The Journal of Technology, Learning, and Assessment* 1 (1), 3–23.

Bocij, P and Greasley, A (1999) Can computer-based testing achieve quality and efficiency in assessment? *International Journal of Educational Technology* 1 (1), 1-21.

Chambers, L (2008) Computer-based and paper-based writing assessment: a comparative text analysis, *Research Notes* 34, 9–15.

Chapelle, C A and Douglas, D (2006) *Assessing Language Through Computer Technology*, Cambridge: Cambridge University Press.

Chodorow, M and Burstein, J (2004) Beyond essay length: evaluating e-rater's performance on TOEFL essays, *TOEFL Research Reports* 73, available online: www.ets.org/Media/Research/pdf/RR-04-04.pdf

Clariana, R and Wallace, P (2002) Paper-based versus computer-based assessment: key factors associated with the test mode effect, *British Journal of Educational Technology* 33 (5), 593–602.

Ellis, R (1994) *The Study of Second Language Acquisition*, 1st edn., Oxford: Oxford University Press.

Ferris, D R (2002) *Treatment of Error in Second Language Student Writing*, Michigan: University of Michigan Press.

Ferris, S P (2002) Writing electronically: The effects of computers on traditional writing, *The Journal of Electronic Publishing* 8 (1), available online: quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0008.104 [DOI: http://dx.doi.org/10.3998/3336451.0008.104

Goldberg, A, Russell, M and Cook, A (2003) The effect of computers on student writing: a meta-analysis of studies from 1992 to 2002, *Journal of Technology, Learning and Assessment* 2 (1), 1–51.

Green, A and Maycock, L (2004) Computer-based IELTS and paper-based versions of IELTS, *Research Notes* 18, 3–6.

Halliday, M A K and Hasan, R (1985) *Language, context, and text: Aspects of Language in a Social-semiotic Perspective*, Victoria: Deakin University Press.

Hasan, R (1985) The texture of a text, in Halliday, M A K and Hasan, R, *Language, context, and text: Aspects of Language in a social-semiotic perspective*, Victoria: Deakin University Press, 70–96.

Horkay, N, Bennett, R E, Allen, N, Kaplan, B and Yan, F (2006) Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP, *The Journal of Technology, Learning, and Assessment* 5 (2), 3–43.

Lee, H K (2004) A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test, *Assessing Writing* 9, 4–26.

Lee, Y (2002) A comparison of composing processes and written products in timed-essay tests across paper-and-pencil and computer modes, *Assessing Writing* 8, 135–257.

Li, J (2006) The mediation of technology in ESL writing and its implications for writing assessment, *Assessing Writing* 11, 5–21.

Liu, M, Moore, Z, Graham, L and Lee, S (2002) A look at the research on computer-based technology use in second language learning: review of literature from 1990–2000, *Journal of Research on Technology in Education* 34 (3), 1–54.

MacCann, R, Eastment, B and Pickering, S (2002) Responding to free response examination questions: computer versus pen and paper, *British Journal of Educational Technology* 33 (2), 173–188.

McDonald, A S (2002) The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments, *Computers and Education* 39, 299–312.

Neuman, G and Baydoun, R (1998) Computerization of paper-and-pencil tests: when are they equivalent? *Applied Psychological Measurement* 22 (1), 71–83.

Peak, P (2005) Recent trends in comparability studies, *Pearson Educational Measurement Research Reports*, available online: www.pearsonassessments.com/NR/rdonlyres/5FC04F5A-E79D-45FE-8484-07AACAE2DA75/0/TrendsCompStudies_rr0505.pdf

Russell, M (1999) Testing on computers: a follow-up study comparing performance on computer and on paper, *Education Policy Analysis Archives* 7 (20), available online: epaa.asu.edu/epaa/v7n20/

Russell, M and Haney, W (1997) Testing writing on computers: an experiment comparing student performance on tests conducted via computer and via paper-and-pencil, *Education Policy Analysis Archives* 5 (3), available online: epaa.asu.edu/epaa/v5n3.html

Russell, M and Haney, W (2000) Bridging the gap between testing and technology in schools, *Education Policy Analysis Archives* 8 (19), available online: epaa.asu.edu/epaa/v8n19.html

Russell, M and Plati, T (2002) Does it matter with what I write? Comparing performance on paper, computer and portable writing devices, *Current Issues in Education* 5 (4), available online: cie.ed.asu.edu/volume5/number4/

Shaw, S K (2003) Legibility and the rating of second-language writing: the effect on examiners when assessing handwritten and word-processed scripts, *Research Notes* 11, 7–10.

Wang, T and Kolen, M J (2001) Evaluating comparability in computerized adaptive testing: issues, criteria and an example, *Journal of Educational Measurement* 38 (1), 19–49.

Wolfe, E W and Manalo, J R (2004) Composition medium comparability in a direct writing assessment of non-native English speakers, *Language, Learning, and Technology* 8 (1), 53–65.

Wolfe, E W, Bolton, S, Feltovich, B and Niday, D M (1996) The influence of student experience with word processors on the quality of essays written for a direct writing assessment, *Assessing Writing* 3 (2), 123–147.

Test taker familiarity and Speaking test performance: Does it make a difference?

LUCY CHAMBERS RESEARCH DIVISION, CAMBRIDGE ASSESSMENT
EVELINA GALACZI RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL
SUE GILBERT CAMBRIDGE ESOL PROFESSIONAL SUPPORT LEADER, SWITZERLAND

Introduction

This study aims to investigate the effect of candidate familiarity on performance in paired face-to-face Speaking tests. We wish to see if candidates who know each other perform differently on the test from those who do not. Results from this research can help inform test administration procedures at Cambridge ESOL to ensure that the test situation is as fair as possible to all candidates.

In addition to its practical implications for the administration of Cambridge ESOL Speaking tests, this investigation also aims to contribute to the debate on the 'interlocutor effect' (O'Sullivan 2002), by considering how the pairing of familiar/non-familiar candidates could impact on their performance and affect their assessment. The term 'interlocutor effect', i.e. the influence which interlocutors can exert on the discourse produced in a Speaking test and

scores received, covers several interlocutor parameters such as acquaintanceship, gender, age, cultural background, proficiency level, personality and conversational style. The role of the interlocutor effect becomes fundamental in speaking tests which are based on a construct defined in interactional socio-cognitive terms, as is the case with the Cambridge General English tests. An obvious question is how influential these variables are in test discourse.

A growing body of literature has investigated the effect of interlocutor variables in solo, paired and group Speaking tests (see for example, O'Loughlin 2002 and Brown and McNamara 2004 on the effect of gender; Berwick and Ross 1996 and Young and Halleck 1998 on the effect of cultural background; Davis 2009, Nakatsuhara 2009, Norton 2005 on the effect of proficiency level; Berry 1993 and Nakatsuhara 2009 on the effect of personality; Iwashita 1996, Katona 1998, O'Sullivan 2002 on the effect of familiarity). The findings from this large body of literature have often been mixed, with some studies indicating a significant effect of the interlocutor variable under investigation, and others finding only a limited effect for the same variable.

In the context of familiarity, Katona (1998) looked at negotiation of meaning in solo Speaking tests between Hungarian interviewers and test takers. In cases when the interviewer was known to the test taker, the author reported that the negotiation of meaning sequences were more natural, whereas in cases involving an unfamiliar interlocutor, there was a more formal, stilted interaction. The finding could have been confounded, however, by the fact that the study was based on two different conditions – the test takers took a practice test with a known teacher and a live test with an unknown examiner. In another investigation of the effect of test taker familiarity, this time in a paired test, O'Sullivan (2002) focused on the performance of Japanese test takers and found evidence of an acquaintanceship effect, reporting that interviewees achieved significantly higher scores when working with a friend. Importantly, however, the author speculated that the effect of interlocutor familiarity may be culturally specific. This contention echoed an earlier study by Porter (1991), who found no evidence in support of a familiarity effect with Arab test takers.

The available empirical research on the familiarity effect has suggested, therefore, that familiarity with one's test partner could play a role. This finding is perhaps not surprising, as it echoes the socio-linguistic literature which indicates that one factor affecting the way we talk is familiarity between interlocutors (Brown and Levinson 1987, Wolfson 1989). Crucially, however, the available research has also indicated that the role of familiarity in test performance is part of a complex interaction with other background variables. Arguing along similar lines in the context of gender-related effects, Brown and McNamara (2004:533) cautioned against 'any simple deterministic idea that gender categories will have a direct and predictable impact on test processes and test outcomes'. The authors further noted that these variables 'compete in the context of an individual's social identity' and no linear, clear-cut behaviours based on background characteristics can be claimed.

The role of background variables (including candidate familiarity) needs to be considered not just in an empirical, but also a theoretical light in the context of the construct underlying the test. As noted earlier, paired Speaking tests are defined in terms of an interactional construct. Swain (cited in Fox 2004:240) convincingly argues that variability related to different characteristics of conversational partners is what 'happen[s] in the real world. And so they [these variabilities] are things that we should be interested in testing.' She further contends that eliminating all variability in speaking assessment is 'washing out . . . variability which is what human nature and language is all about' (ibid). Coping successfully with such real-life interaction demands, therefore, becomes part of the construct of interactional competence which underlies paired Speaking tests. This does not imply, however, that test developers should disregard the effect of interlocutor variability. Test providers have an ethical responsibility to construct tests which are fair and do not provide (intentionally or unintentionally) differential and unequal treatment of candidates. One such approach to controlling some of the variability introduced in a paired test is through the use of multiple response formats in the test, which generate different types of talk, e.g., test taker/examiner, test taker/test taker, and test taker only. The use of multiple response formats in Speaking tests (which is the case in all Cambridge English paired Speaking tests) reduces the inherent unpredictability in paired tests as it allows different types of talk to be generated, based on different levels of test control. Different response formats, therefore, have the potential to optimise the benefits of the paired format, while controlling for possible caveats associated with interlocutor variables.

In addition to considering the familiarity effect in an empirical and theoretical context, it is also important to address it from a stakeholder point of view. A multiple-method approach which incorporates both statistical score data and qualitative data based on test taker perceptions would bring together essential points of view and contribute to a more comprehensive understanding of the role of this variable. Consultation with stakeholders is an important element of the test development and validation process at Cambridge ESOL. In fact, the initial driving force for the present study was a query from an exam centre regarding the role of test taker familiarity and a request for further investigation. The present study, therefore, draws on a range of data-collection sources: quantitative test taker score and questionnaire data, as well as qualitative candidate interview data and uses a convergent parallel mixed-method design where qualitative and quantitative data are related and interpreted together in the investigation of the research questions. We hope that such a mixed-method approach will provide useful insights for test developers and test users about the role of test taker familiarity, so that its effect is better understood and more clearly conceptualised.

Methodology

Research questions

1. What is the effect of test taker familiarity on the *scores* awarded in a *Cambridge English: First (FCE)* paired test?
2. What are *test taker perceptions* about the effect of familiarity on their performance?

Context of the study

A total of 641 candidates taking the *Cambridge English: First* Speaking test in three Cambridge ESOL examination centres in Switzerland – Bern (German L1), Geneva (French L1) and Ticino (Italian L1) – in the 2011 summer session were involved in various stages of the project. An important premise of the study was the need to target a range of cultural contexts, in order to reduce as much as possible any potential bias introduced by the cultural setting.

Before proceeding further, it's important to note a methodological limitation of the study, namely the fact that mostly European test takers representing a narrow range of L1s were involved in this study. As we know from the research cited in the Introduction, cultural background could potentially influence test taker performance, as it works alongside other background variables such as gender and familiarity. The findings, therefore, have to be considered within the context underlying the study. It is hoped that the research questions of interest here can be explored in a range of different cultural and L1 settings, so that more generalisable results can be produced.

Data collection and analysis

Prior to the start of data collection, exam centres, schools and candidates were informed via a letter distributed through the centre of the nature of the project and what was involved. The *Cambridge English: First* exam was chosen as it is one of the most widely taken Cambridge English exams. It also has a paired Speaking test with a candidate-candidate interaction task (Part 3) which involves the candidates working together to discuss a topic and make a decision. If candidate familiarity were to play a role in test performance, it is likely that it would be most evident in this part.

For the remainder of the report candidates who know each other will be referred to as *familiar* and those that do not know each other as *non-familiar*.

The data collection involved several stages and sources:

Questionnaires

All candidates sitting a *Cambridge English: First* Speaking test completed a short 5-minute questionnaire prior to their tests. The questionnaire sought to establish whether or not, and how well, candidates knew their Speaking test partner. A frequency analysis of the questionnaire responses was carried out. In total there were 629 respondents (98% response rate).

Candidate exam marks

The questionnaires included name, centre number and candidate number so that they could be linked to exam score data, allowing for comparisons to be made between the scores of the familiar and non-familiar groups.

Candidate interviews

Four familiar pairs and four non-familiar pairs from each centre were invited to take part in a feedback session directly after their Speaking test in order to gather their perceptions about the test. Invitees were given the option not to be involved. At each exam centre, the Centre Exams Manager organised and selected the interviewees. In order to plan for the feedback sessions, certain assumptions about the respondents' level of familiarity had to be made. Candidates from the same school were treated as familiar; individual candidates (i.e. not from a test preparatory school) were treated as non-familiar.

Each candidate was interviewed alone, with their partner being interviewed concurrently. These sessions were conducted by an interviewer who was not one of the candidate's examiners and were in the main L1 for each centre. All interviews were audio-recorded and protocol forms were completed by the interviewer. Data from the feedback interviews was analysed to establish how the candidates felt about knowing or not knowing their partner and whether they thought this had any impact on their performance.

Table 1 shows the actual number of feedback interviews. Due to practicality issues on the test day, one centre was unable to get enough non-familiar candidates. Another centre recruited extra familiar and non-familiar candidates.

Table 1: Number of interviewed candidates by centre and familiarity

Location	Geneva	Bern	Ticino	Total
Familiar	8	8	14	30
Non-familiar	5*	8	10	23
Total	13	16	24	53

*Odd number due to a trio-format test (this happens if there is an odd number of candidates at a centre).

Audio-recording of tests

The candidates involved in the feedback sessions had their Speaking tests recorded. In addition, approximately one day of tests was audio-recorded at each centre; these contained a mixture of familiar and non-familiar pairs. This data can be analysed to further explore trends/findings from the current study. A discourse analysis of the candidate/candidate interaction will be carried out at a later stage and is beyond the scope of the current study.

Study participants

Before proceeding to a discussion of the findings, it is important to consider some of the candidate background information and its distribution in the familiar/non-familiar groups in more detail, since any large difference in background variables between these two groups may confound the analysis. If the majority of familiar candidates were female, for example, and the non-familiar ones were male, this could play a role in their test performance, since gender has been shown to potentially play a role in Speaking tests (Brown and McNamara 2004). As will be seen, both familiar and non-familiar groups were generally comparable in terms of participants' background variables.

Table 2 shows the breakdown of the familiar/non-familiar groups overall and by centre. Out of the 629 candidates who completed the questionnaire, 56% were part of a familiar pair and 43% were in an unfamiliar pair. The majority of candidates came from the Bern centre.

In terms of L1, of the majority of candidates matched the L1 of each test centre (Geneva – 58% French, Bern – 58% German and Ticino – 79% Italian). In terms of gender, 60% of the study participants were female and 40% male. The

Table 2: Questionnaire sample by centre and familiarity

	Geneva (French L1)		Bern (German L1)		Ticino (Italian L1)		Total	
	N	%	N	%	N	%	N	%
Familiar	65	62%	262	57%	25	40%	352	56%
Non-familiar	37	36%	195	42%	37	60%	269	43%
Blank	2	2%	6	1%	0	0%	8	1%
Total	104	100%	463	100%	62	100%	629	100%

Table 3: Comparison of test means of familiar/unfamiliar candidates

	Non-familiar (N = 269 candidates)		Familiar (N = 352 candidates)		
Test paper	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Difference in means**
Speaking	29.75	4.80	31.15	4.56	−1.39
Reading	29.78	6.66	30.67	6.17	−0.88
Writing	29.62	3.30	29.98	3.19	−0.37
Use of English	24.49	6.47	25.23	6.04	−0.74
Listening	28.23	6.93	30.00	6.22	−1.77

gender distribution within the familiar/non-familiar groups was comparable.

Approximately two thirds of candidates were between 17 and 25 years of age (65%), 12% were aged 16 or under and 24% were aged 26 or over. The familiar group participants were slightly younger than their non-familiar counterparts. This trend is explained by the organisation of Speaking tests in Switzerland, where groups of candidates from each school come to the centre in a block for their test, thus it is expected that the majority of school age candidates would know each other. Older candidates tend to enter the exam as individuals and so are likely to be in non-familiar pairs. We believe this different age distribution to be a minimal limitation of the study, since the difference in age was very small.

For those candidates who knew their partner, 85% knew each other from school and 12% from private language class. The remainder knew each other from work, social activities, church or were related.

Results and discussion

Familiar vs. non-familiar candidate performance

Mean scores for the familiar and non-familiar groups were compared for the whole exam, including speaking (Table 3). The results indicate that taken as a whole the familiar candidates scored higher on *all papers* than the non-familiar candidates; they also consistently showed a lower standard deviation. We could hypothesise that the ability of this group was slightly higher and less varied because they had attended similar test preparation courses. The Speaking and Listening papers showed the greatest difference in mean scores, although this was less than 2 points (out of 40). T-tests showed these results to be significant for Speaking ($t(619) = -3.7$, $p = 0.0002$) and Listening ($t(619) = -3.4$, $p = 0.0009$); however tests for effect size[1] showed the effect to be small (0.15 and 0.13 respectively). The fact that differences were found for all test papers suggests that differences in Speaking means between the two groups are likely to be a result of differences in ability rather than due to the effect of candidate familiarity.

In addition to investigating the familiar/non-familiar end of the acquaintanceship continuum, the analysis also focused on: a) differences in performance based on *how well* candidates knew their partner and b) whether they described their partner as a *friend*. In the former analysis, a Likert scale of 1–5 was used in the questionnaire and scores of 1 and 2 were compared to scores of 4 and 5 (non-familiar candidates were given a score of 1). No significant difference was found for any skill. In the latter analysis significant differences were found for both Speaking and Listening. However, effect sizes were small (0.12 for both papers), giving further evidence of there being no meaningful effect of knowing your partner on Speaking mean scores.

The Speaking assessment criteria used to assess speaking were also examined, to explore whether there were any meaningful differences in performance in any of the criteria, but especially in *the Interactive Communication* criterion. *Interactive Communication* refers to the candidate's ability to take an active part in the development of the discourse, to initiate and respond effectively, to develop discussions and maintain the interaction. As this criterion is based on joint contributions, it is the one most likely to be affected by one's partner. Table 4 compares the means of each criterion and as can be seen, the familiar candidates scored slightly higher, as expected, given the total mean scores for Speaking. However, the individual assessment criteria appear to be comparable, and any differences are generally within half a mark (out of 10 in the case of GV, DM, PR, IC, and 20 in the case of GA). *Interactive Communication* does not show any marked difference, further evidence that familiarly does not appear to have an effect on these candidates.

[1] Effect size is a way of quantifying the difference between two groups; it has advantages over the use of tests of statistical significance alone, as it emphasises the size of the difference rather than confounding this with sample size (Coe, 2002).

Table 4: Comparison of criterion means of candidates who know/don't know each other

	Non-familiar (N = 269 candidates)		Familiar (N = 352 candidates)		
Test paper	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Difference in means**
Grammar and Vocabulary (GV)	7.03	1.50	7.36	1.36	−0.33
Discourse Management (DM)	7.43	1.37	7.70	1.37	−0.27
Pronunciation (PR)	7.45	1.24	7.70	1.26	−0.25
Interactive Communication (IC)	7.66	1.29	8.06	1.25	−0.40
Global Achievement (GA)	15.07	2.71	15.90	2.45	−1.83

To sum up, the comparative analysis of the scores of the familiar and non-familiar groups indicated small, but not meaningful differences in overall Speaking test performance and performance by assessment criteria.

Candidate feedback

The feedback sessions started with some general questions about the test itself, which was both to obtain some background information and to ease the participants into the session. In response to the question *Why did you take FCE?,* the most common answer for both groups was to help with their future plans. For many of the familiar group it was compulsory; in contrast many of the non-familiar group cited work and pleasure as additional reasons. Both groups appeared to have done similar preparation for the test and the majority of candidates found the experience enjoyable.

The familiar candidates answered some additional questions to ascertain how well they knew their partner; 13 out of 30 said they knew their partner 'very well', 15 'somewhat well' and two 'not very well'. All members of this group knew each other from school; six had known each other less than one year, six: 1–2 years, 16: 2–3 years and two for more than three years. Thirty-one classed their partner as a friend and nine as a classmate; all said they got on well with their partner. For the non-familiar group, 14 out of 23 spoke briefly to their partner before the test, nine did not.

The next set of questions attempted to explore how the candidates felt about their performance and whether they felt their partner influenced it in any way. A summary of the responses is detailed under each question – only factors mentioned by three or more of the participants are mentioned.

Do you think that you performed well in the Speaking test?

The majority of candidates in both groups felt they performed well in the test, some felt they did ok and a minority weren't sure (see Table 5). No one felt they had done badly.

Table 5: Candidate perception of performance

	Familiar	Non-familiar
Good	14 (47%)	13 (57%)
Ok	10 (33%)	7 (30%)
Unsure/neutral	6 (20%)	3 (13%)
Total	30	23

What do you think influenced the way you performed?

The main factors the familiar group cited were: test anxiety, knowing their partner, exam preparation and the examiner. The non-familiar group cited the fact that it was an exam, test anxiety, getting on well with their partner and the examiner. It is interesting that both groups mentioned their partner, though more people in the familiar group (12) mentioned this than the non-familiar group (7), though often it was embedded in a list of other factors. This is perhaps not surprising, since the peer interlocutor is an inevitable key variable in a paired Speaking test. It should also be noted that both the interviewers and interviewees knew that the topic under discussion was about partner familiarity so it was possibly forefront in their minds.

Did you enjoy doing a Speaking test with your partner today?

The majority of candidates from both groups enjoyed doing the test with their partner (familiar 27, non-familiar 20), as can be seen in Table 6. Sixteen of both the familiar and non-familiar candidates cited their partner as a reason they found the test enjoyable (e.g. the partner could help them through providing ideas for the discussion). Only one familiar and two non-familiar candidates were neutral about their experience. The familiar candidate would have preferred a partner they didn't know and one of the non-familiar candidates cited that they didn't enjoy the experience because it was a test. No one was negative.

Table 6: Candidate expression of enjoying doing the test with their partner

	Familiar	Non-familiar
Enjoyed	27 (90%)	20 (87%)
Unsure/neutral	1 (3%)	2 (9%)
Blank	2 (7%)	1 (4%)
Total	30	23

Representative comments[2] below:

- *More relaxing like this. Felt on the same wavelength as my partner because I knew her.* (familiar)
- *Because she has lots of imagination, she can help you. If you don't know what to say, sometimes she can help.* (familiar)
- *Because we got to know each other better while preparing the test together.* (familiar)
- *Especially in Part 3, which went well because they developed the discussion together, they understood each other well.* (familiar)
- *She was prepared and in Part 3 there was a good interaction.* (non-familiar)
- *He was more relaxed than I was, helped me, didn't know him so felt no shame if I said something wrong or stupid.* (non-familiar)

[2] Note: the use of 1st/3rd person is dependent on the style with which the interviewers recorded their protocol summaries.

– *But it would have been the same with someone else. It's not that important who the partner is.* (non-familiar)

– *Because you could get to know another person and she was interested in what she was saying.* (non-familiar)

Did you feel you worked well together?

The majority of candidates in both groups felt they worked well with their partner (see Table 7); reasons cited were quality of interaction and team work:

– *The interview was balanced, each spoke for the same amount of time and helped each other when necessary.* (familiar)

– *We 'connected'.* (non-familiar)

– *There was a good rapport between us.* (non-familiar)

The candidate that selected 'not always' felt that the interaction in Part 3 (the candidate/candidate interaction task) wasn't always easy; although in general she enjoyed doing the Speaking test with her partner.

Table 7: Candidate perception of working well together

	Familiar	Non-familiar*
Worked well	28 (93%)	20 (87%)
Unsure/neutral	1 (3%)	1 (4%)
Not always	0	1 (4%)
Blank	1 (3%)	1 (4%)
Total	30	23

*Percentages do not sum to 100 due to rounding.

Do you feel that your partner helped you or made it a bit harder for you?

Generally, familiar candidates felt that their partner helped them; non-familiar candidates were either neutral or positive about their partner's influence (see Table 8). Familiar candidates cited that their partner made them feel relaxed and they had good team work (*We gave each other opportunities to speak, Her partner helped her to be more calm*); non-familiar candidates also cited the positive influence of their partner (*In Part 3 we helped each other. In Part 1 and 2 it doesn't matter, He organised the conversation, switched to the next picture and was co-operative).*

Table 8: Candidate perception of partner influence

	Familiar	Non-familiar*
Partner helped	24 (80%)	7 (30%)
Neutral	5 (17%)	10 (43%)
Partner made it harder	0	3 (13%)
Partner unimportant	1 (3%)	2 (9%)
Blank	0	1 (4%)
Total	30	23

*Percentages do not sum to 100 due to rounding.

Did you have the opportunity to say everything that you wanted in the test?

A large number of candidates from both groups felt they did not have adequate opportunity to say all they wanted in the Speaking test (see Table 9); the main reason cited was not enough time (*The examiners stop you from saying more not the partner* (non-familiar*), Sometimes the time allocated was too short, couldn't finish my sentence* (familiar). Importantly, none of the candidates cited their partner as a cause.

Table 9: Candidate perception of opportunity

	Familiar	Non-familiar
Opportunity to say everything	13 (43%)	12 (52%)
Not enough opportunity	15 (50%)	11 (48%)
Blank	2 (7%)	0
Total	30	23

Did your partner give you the opportunity to say everything that you wanted?

Generally candidates felt that their partner gave them adequate opportunity to speak (see Table 10). The main reason cited was collaboration:

– *One spoke and then invited the other to contribute.* (familiar)

– *Sometimes they both started to speak at the same time; when this happened, sometimes she let her partner speak, sometimes her partner let her speak. When it happened, they stopped and looked at each other – it was really nice!* (familiar)

– *He let me speak, he supported ideas when I moved to a new picture.* (non-familiar)

– *We encouraged each other, the interaction in Part 3 was good.* (non-familiar)

The three familiar candidates who disagreed with the statement said that their partner spoke too much or interrupted them (*And even in Part 4 when the question was for me, my partner intervened, she interrupted me, it was a 'competition' between us*).

Table 10: Candidate perception of opportunity given by partner

	Familiar	Non-familiar
Opportunity to say everything	23 (77%)	20 (87%)
Not enough opportunity	3 (10%)	0
Blank	4 (13%)	3 (13%)
Total	30	23

Do you think if your partner was someone else (friend/stranger) it would have made a difference to how you performed in the test or in specific parts of the test?

The majority of familiar candidates in the study felt they preferred being partnered with someone they know; this was largely due to knowing what to expect, i.e. their partner's

Table 11: Candidate perception of having a different partner

	Familiar	Non-familiar*
Better with friend	13 (43%)	1 (4%)
Better with stranger	3 (10%)	7 (30%)
No difference	5 (17%)	6 (26%)
Depends/different/don't know	9 (30%)	8 (35%)
Blank	0	1 (4%)
Total	30	23

*Percentages do not sum to 100 due to rounding

language level, pronunciation and personality (see Table 11). There was a perception that they would have been more nervous with someone they didn't know. In contrast, the non-familiar candidates did not see a benefit in being with a friend, but showed similar concerns about a potential partner's personality and language level. It would appear that it is the quality of the *partnership match* (i.e. how well balanced the interlocutors are in terms of language level, personality variables etc.) that is important rather than knowing the partner per se.

Some comments from familiar candidates:

- *The partner doesn't have to be a friend but what is important is their level is more or less the same and that the preparation for the exam is the same.*
- *It worked well with her partner because they were about the same level. It could have been a problem if the person had been much better or selfish. A weaker candidate wouldn't have been good – it would depend on the social skills of the partner. That would have been more important than whether or not the partner was a schoolmate.*
- *It could have been OK with an unknown partner, but you wouldn't know the level, at least at the start and there would have been more fear about Part 3, in case she didn't understand.*

Comments from non-familiar candidates:

- *Because they had more or less the same level, so she felt more relaxed. The important thing was that the level of English was similar. She would have felt panicky if the level was very different. She felt it was better NOT to take the test with a friend, because you're used to speaking your own language with your friends, so it would strange to speak in another language. Maybe you're also influenced by your friends – there is also fear of judgement of your friends.*
- *I feel less at ease with friends, because I worry about them judging my performance. But interrupting friends is easier.*
- *With friends, the competition element comes into play. But this can also push us to do better in a test situation.*

Conclusions and recommendations

1. What is the effect of test taker familiarity on the scores awarded in a *Cambridge English: First* paired test?

As noted above, the comparative analysis of the scores awarded to the familiar and non-familiar groups in the Swiss centres of interest indicated small, but not meaningful, differences in overall Speaking test performance and performance by assessment criteria.

2. What are test taker perceptions about the effect of familiarity on their performance?

The questionnaire and interview candidate feedback indicated that the Swiss candidates in this study did not perceive familiarity with their partner as affecting their performance. In other words, taking the *Cambridge English: First* Speaking test with a friend was not perceived – by either the familiar or non-familiar group – as giving an unfair advantage. However, the candidates felt very strongly about the effect of the test partner's language ability on their test performance.

It is worth noting that the appropriateness of proficiency matching is controlled for in the Cambridge English paired tests, since they are divided into language levels (e.g. *Cambridge English: First* is aimed at CEFR B2 level). Such fine-tuned targeting of the level of the exam avoids a significant proficiency mismatch between the paired candidates. Other examinations which are aimed at a much wider proficiency range, such as *IELTS*, do not use a paired format because of the possibility for a mismatch in proficiency level of candidates. The test taker's concern with an adequate match in test taker proficiency level also provides valuable insights into a possible replication of the study: it would be useful to replicate this study at different proficiency levels, in order to explore the interaction between the role of test taker familiarity and proficiency level. A higher proficiency level could give more linguistic confidence and mitigate any potential familiarity effect. In contrast, a lower level proficiency could be more affected by interlocutor variables such as familiarity, which would have implications for the use of open-ended paired tasks at lower levels.

To sum up, this study has revealed some important insights about candidate performance and candidate perceptions about their performance in the context of candidate familiarity. Both the statistical analysis of candidate score data and the qualitative analysis of questionnaire and interview responses have indicated that candidate familiarity plays a minimal role. It is important to remember, however, that this study was done in a European context and so any generalisations about the results need to be supported by data gathered from a range of different cultural settings.

It is also necessary to keep in mind the complexity of investigating the issue of interlocutor variables, which are difficult to isolate and to comprehensively conceptualise (e.g. even the way 'familiarity' is defined is complex, due to its gradations). As such, test developers face a fine balancing act of reconciling research findings, theoretical conceptualisations and stakeholder perceptions. It is hoped that the present research study has provided useful insights from all these three perspectives which would support test developers and test users.

References and further reading

Berry, V (1993) Personality characteristics as a potential source of language test bias, in Huhta, A, Sajavaara, K and Takala, S (Eds) *Language Testing: New Openings*, Jyvaskyla, Finland: Institute for Educational Research, University of Jyvaskyla, 115–124.

Berwick, R and Ross, S (1996) Cross-cultural pragmatics in oral proficiency strategies, in Milanovic, M and Saville, N (Eds) *Performance Testing, Cognition and Assessment. Selected Papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem*, Studies in Language Testing volume 3, Cambridge: UCLES/ Cambridge University Press, 34–54.

Brown, A and McNamara, T (2004) The devil is in the detail: Researching gender issues in language assessment, *TESOL Quarterly* 38 (3), 524–538.

Brown, P and Levinson, S C (1987) *Politeness. Some Universals in Language Usage*, Cambridge: Cambridge University Press.

Coe, R (2002) *It's the Effect Size, Stupid. What effect size is and why it is important*, paper presented at the Annual Conference of the British

Educational Research Association, University of Exeter, England, 12–14 September 2002, available online: www.leeds.ac.uk/educol/documents/00002182.htm

Davis, L (2009) The influence of interlocutor proficiency in a paired oral assessment, *Language Testing* 26 (3), 367–396.

Fox, J (2004) Biasing for the best in language testing and learning: An interview with Merrill Swain, *Language Assessment Quarterly* 1 (4), 235–251.

Iwashita, N (1996) The validity of the paired interview format in oral performance assessment, *Melbourne Papers in Language Testing* 5 (2), 51–65.

Katona, L (1998) Meaning negotiation in the Hungarian oral proficiency examination of English, in Young, R and He, A (Eds) *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency*, Amsterdam/Philadelphia: John Benjamins, 239–267.

Nakatsuhara, F (2009) *Conversational Styles in Group Oral Tests: How Is The Conversation Constructed?*, unpublished PhD thesis, University of Essex.

Norton, J (2005) The paired format in the Cambridge speaking tests, *ELT Journal* 59 (4), 287–297.

O'Loughlin, K (2002) The impact of gender in oral proficiency testing, *Language Testing* 19 (2), 169–192.

O'Sullivan, B (2002) Learner acquaintanceship and oral proficiency pair-task performance, *Language Testing* 19 (3), 277–295.

Porter, D (1991) Affective factors in language testing, in Alderson, C and North, B (Eds) *Language Testing in the 1990s*, London: MacMillan, 32–40.

Wolfson, N (1989) *Perspectives: Sociolinguistics and TESOL*, Boston: Heinle & Heinle.

Young, R and Halleck, G B (1998) 'Let them eat cake!": Or how to avoid losing your head in cross-cultural conversations, in Young, R and He, A (Eds) *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency*, Amsterdam/Philadelphia: John Benjamins, 359–388.

A reading model for foundation year students at a tertiary institution in the United Arab Emirates

HELEN DONAGHUE CENTRE FOR TEACHING AND LEARNING CO-ORDINATOR, SHARJAH HIGHER COLLEGES OF TECHNOLOGY, UNITED ARAB EMIRATES

JASON THOMPSON ENGLISH TEACHER, SHARJAH HIGHER COLLEGES OF TECHNOLOGY, UNITED ARAB EMIRATES

Introduction

This article outlines the implementation of a reading model for foundation year students at a tertiary institution in the United Arab Emirates (UAE). The aim of the reading model was to uncover and address difficulties students experienced in second language (L2) reading and help them achieve a reading level to enable progression to a bachelor degree course. We started by designing a reading construct, based on a socio-cognitive model of language proficiency (Weir 2005), which formed the basis for writing learning outcomes. We then initiated a cycle of curriculum planning, teaching, test planning, testing, analysis and feedback. In this article we report on the design process and implementation of the reading model.

Reading skills

Researchers are still interested in investigating the nature of L2 reading and how people read. Although different approaches and theories have been suggested, a consensus has yet to be reached. Different lists or taxonomies have been developed from Munby (1978) to Nuttall (1996) which break down the skill of reading into sub-skills. This approach remains popular and many English as a Foreign Language (EFL) course books, teacher education courses and tests adhere to the notion that sub-skills exist, despite more current thinking questioning their validity:

> . . . their origins are more frequently in the comfort of the theorist's armchair than they are the result of empirical observation . . . they are frequently ill defined (or undefined) and give a misleading impression of being discrete when in fact they overlap enormously (Alderson 2000:11).
>
> The [sub-skills] approach has mainly been based on informed intuition rather than empirical research but has been found useful by a generation of teachers (Khalifa and Weir 2009:39).

More recently, researchers mostly agree that the reading process combines 'bottom-up' processing, such as word recognition and decoding and 'top-down' schemata. Khalifa and Weir (2009) propose a socio-cognitive framework for reading which identifies different types of reading and the cognitive processes they elicit. We decided to use this framework as a basis for our reading model because it encompasses the mental processes involved in reading and views the use of language tasks as social rather than purely linguistic phenomena.

Investigating the reading skills of a group of students in a particular context is perhaps more useful than large-scale, generalisable research into L2 reading. There are a number of factors which make universal research difficult. L2 students are often transient, may study for short periods and, perhaps most importantly, the complex nature of reading introduces many variables. L2 learners can have different socio-cultural and educational backgrounds, varying degrees of first language reading skills and strategies, different levels of L2 proficiency and different writing scripts. Learners can have differing degrees of background knowledge, interests and

motivation. There is variation in text types, text complexity, topics, reasons for reading and ways of reading: all this makes wide-scale research a challenge. Thus, action research in our own contexts is perhaps a more appropriate method of investigating reading skills:

> We should not wait for sweeping assertions from research, nor should we be swayed by claims of 'perfect' classroom solutions. Rather we should use our own classrooms, and our own students, as a forum for meaningful classroom-based research (Grabe and Stoller 2000:62).

Context and problems

Although instruction at our college is in English, Emirati students' first language is Arabic and most entrants do not meet the required *IELTS* Academic band 5 for entry to bachelor programmes. Thus, depending on their level of proficiency, most students spend from six months to two years studying on a foundation course which includes 20 hours of English instruction a week, to help them achieve the required English level. Test score evidence (both institution-based tests and *IELTS* scores) indicate that reading is consistently the students' weakest skill, which is consistent with national performance – for the past three years, the UAE average *IELTS* Reading band has been at least one band below the international mean (e.g. In 2010, the international mean in reading was 5.97 and the UAE mean was 4.8). Our students also report that little attention is paid to reading in English at their schools. We conducted interviews with 15 randomly chosen students and they indicated that little emphasis was given to reading skills, as these comments illustrate:

> At school they didn't focus on the skills we need for reading. They just give us the text to read and then we answer the questions, and then they give us the answers to the questions. We didn't benefit from this (Interviewee 4).
>
> [The school teacher] didn't explain us about the task or how can we do this task. We didn't get better at reading because she didn't give us rules or ways to do it. She only give us task – ok do it. How you can do, do it (Interviewee 11).
>
> In the school they didn't learn us how to make the reading to answer but only when we have exam they give us the passage and only the question true and false and some question about the passage and multiple choice but they didn't tell us how to answer these questions. We don't have reading classes, only the book. In the unit there was things to read but only we have to learn the vocab and a lot of grammar (Interviewee 6).

L1 reading ability should also be considered. If students cannot read well in their first language, L2 reading will be difficult (Alderson 2000:38). Educational reports suggest poor L1 literacy in the UAE where Emirati students score well below other developed and developing countries on the Organisation for Economic Co-operation and Development (OECD) scale:

- only 0.5% of Emirati students attain high reading skills (Knowledge and Human Development Authority 2009), and almost 80% of students at government schools in Dubai have reading skills below a student of the same age in the lowest 25% of the distribution at private UK and International Baccalaureate (IB) schools (Knowledge and Human Development Authority 2009)
- 47% of Emirati students have 25 or fewer books at home (Knowledge and Human Development Authority 2007).

Thus, many students may enter the college with poor reading skills and little experience or success in reading in either Arabic or English.

In addition, prior to our intervention, the foundation reading course was, in our opinion, poorly planned and, more importantly, the learning outcomes for the course were poorly defined. Learning outcomes are usually statements of what a student is expected to know, understand or be able to do after the process of learning and how that learning will be demonstrated (Gosling and Moon 2001) but the original outcomes from the reading course often deviated from this definition. For example, this learning outcome from the original reading course:

- *Students will develop a habit of reading.*

does not refer to knowledge, understanding or ability and, in addition, would be difficult to demonstrate, as would:

- *Students will identify and use prior knowledge of a topic to support reading.*

Some were not outcomes but rather a description of test tasks, e.g.:

- *Students will respond accurately to closed-response comprehension questions about gist, inference, and detail.*

while others lacked clarity, e.g.:

- *Students will apply strategies to locate specific information in a range of linear and non-linear texts by the end of the course.*
- *Students will use reading to find specific information.*

Some learning outcomes contained more than one operation, e.g.:

- *Students will employ skimming and scanning techniques to quickly locate specific facts in linear and non-linear texts in a limited amount of time by the end of the course.*

As well as outcomes being ill formulated, they were not comprehensive. Five learning outcomes described the skill of locating specific information but other reading skills such as identifying main ideas/supporting details or making inferences were not covered. A course based on these learning outcomes, we felt, was unlikely to improve students' reading ability, hence our intervention.

Intervention

A reading construct

We started by trying to define a reading construct which we could then use for teaching and testing purposes. A construct is a theoretical definition of the underlying abilities of the skill we are hoping to develop: essentially, what we want to teach and measure. However, developing an idea of what constitutes reading has so far been an elusive goal. Despite this, it is important to try to develop an understanding of what we want to teach and test, even though we may know in advance that our understanding of the construct of reading is 'faulty, partial and possibly never perfectible' (Alderson 2000:2). The tests we devise based on this construct may be imperfect, but the

process of teaching, testing and analysis may lead to a better understanding of our students' ability:

> . . . it is only by trying to operationalise our theories and our understandings of the constructs through our assessment instruments that we can explore and develop our understanding (Alderson 2005:2).

Thus, the purpose of the reading construct was to form the basis of a reading curriculum and also inform test writers and so help ensure alignment between teaching and testing. The first step we took towards attempting to define a reading construct for our foundation course was to consider the Common European Framework of Reference for languages (CEFR) (Council of Europe 2001).

The CEFR

The CEFR has a number of strengths. It details functional competencies across six levels, realised in can do statements. The competencies are well mapped out and communicative language activities and strategies are comprehensively listed. The CEFR is flexible and not prescriptive so can be adapted to suit different contexts and purposes. It also has a flexible branching approach within which levels can be broken down into smaller units to finer distinguish between levels. The framework of levels helps with planning a syllabus and can take into account progression and enable continuity between different levels. The framework is also useful for the selection of course materials as many course books are mapped to the CEFR (for example, *Total English* (Pearson), *English Unlimited* (CUP), *New Headway* and *New English File* (OUP)). The CEFR levels are aligned with the Cambridge ESOL exams so these can be referenced in curriculum planning and test writing. Developing courses and designing or selecting materials is further assisted by themes outlined for each level. The use of strategies is emphasised in the CEFR manual:

> The use of communication strategies can be seen as the application of the metacognitive principles: pre-planning, execution, monitoring, and repair action to the different kinds of communicative activity: reception, interaction, production and mediation (Council of Europe 2001:57).

We wanted to make sure that this was part of our teaching so that students were not only taught skills, but also clear strategies that would help them to complete a task successfully.

However, the CEFR also has limitations. Fulcher suggests that because the scales are not based on primary data, descriptors may be inadequate and levels flawed (2010:116). Indeed, similar descriptions occur at different levels, some operations (e.g. recognise, infer) appear in some levels but not others and many of the terms e.g. '*simple*' are not operationalised (Alderson, Figueras, Kuijper, Nold, Takala and Tardieu 2004). More importantly, the CEFR lacks some contextual features that we wanted to include such as task type and discourse mode: the CEFR gives little advice about what types are suitable for different levels and skills. We also added specific text lengths as we felt the CEFR descriptions of '*long*' and '*short*' were too vague.

The CEFR is also limited in terms of cognitive processing (Weir 2005). Depending on the task at hand, readers will use different kinds of reading such as slow, expeditious, global, or local (Khalifa and Weir 2009) but not all reading types are covered at every level, for example scanning is not mentioned until CEFR B1 level. The CEFR also lacks specification of sub-skills such as comprehension of main ideas, recognising the structure of a text or recognising connections between parts of a text and we considered these an important element of our reading construct. Therefore, to further define our reading construct, we used Khalifa and Weir's reading model (2009:43). As it is based on a socio-cognitive model of language proficiency, the model addresses contextual and cognitive parameters missing from the CEFR. This model details the metacognitive activity of the reader, including the goal setter (selecting the appropriate type of reading, text or task), the processing core which involves the levels of cognitive processing involved in building up meaning at word, phrase, sentence and text level and the knowledge required for understanding the text. We thought it important to consider how language develops across the proficiency levels in terms of cognitive processing, so we added both goal setting and aspects of cognitive processing to our construct.

The construct

We wrote a construct for each of the four levels in our foundation course (for example, see Appendix A for the level 3 construct), using the CEFR levels, Khalifa and Weir's (2009) reading model and our knowledge of the level and types of texts we expected our students to be able to deal with. The construct details texts to be used at each level, specifying type of text (e.g. newspaper articles, brochures, websites), discourse mode (e.g. descriptive, narrative, instructive, expository, argumentative), text purpose (e.g. referential, poetic, emotive), nature of information (e.g. concrete, abstract), cognitive demands (e.g. familiar or idiomatic vocabulary, simple or complex grammatical structures), text difficulty (using Kincaid, Fishburne, Rogers and Chissom 1975 as a rough measure) and text length. We then specified types of reading and the cognitive processing associated with these, and established what students should be able to do in terms of reading skills, which gave us learning outcomes, and therefore the basis for the courses at each level. The learning outcomes were then organised by level of processing (sentence to paragraph to text level) and then learning outcomes were mapped into the curriculum so teachers knew which outcomes to cover when. Appendix B is an example of the level 3 learning outcomes.

The cycle

With the construct and learning outcomes in place, we initiated a cycle of teaching, testing, analysis and feedback.

Teaching and testing

The semester was broken down into six-week blocks and at the end of each block students did a progress test which assessed the learning outcomes covered in the previous six weeks. We provided teachers with teaching material and a plan indicating when material should be taught as well as making explicit which learning outcomes should be addressed in class. We wrote three progress tests for each level, starting with test specifications which included text length and difficulty, themes, learning outcomes and task types. We then searched for suitable texts and designed

tasks based on the learning outcomes to be tested. We sent teachers an email at the beginning of every six-week block (and another two weeks before the test) with information about the test specifications so they were clear about which text types, learning outcomes and task types students would encounter in the test.

Analysis and feedback

After doing the tests on paper, students entered their answers on the online learning platform Blackboard Vista which gave us immediate information such as the level of difficulty of the items, an individual student's performance on each item, and all students' answers for each item. Using this information, the next stage in the process was to write a report for teachers, which detailed information such as the percentage of students in their class that got each question correct and common mistakes and errors that teachers needed to be aware of to pass on to students. We also provided analysis by learning outcome, highlighting student difficulty or success for each outcome and suggested remedial work when needed. With the report focusing on the learning outcomes that required further work in the classroom, the cycle came back to teaching.

Evaluation

Eighteen months after we initiated this reading model, we do not, of course, have a comprehensive understanding of the nature of L2 reading but this process of teaching, testing and analysis has provided us with valuable information about our students' reading. We know, for example, that most of our students can scan fairly competently so we now devote less class time to this skill and include fewer items testing it. With inferencing and reading for main ideas, however, students tend to be less successful so we have built in more instruction and help in class with these skills. Teachers have reported that this systematic approach to curriculum and testing, as well as the fact that test analysis informs teaching, has enabled them to better address their students' difficulties in reading.

We interviewed 15 students randomly chosen from different levels and classes to try to find out how they felt about the foundation reading course. All of them showed an awareness of reading skills, strategies and different ways of reading and many reported they felt more positive about reading. For example:

> It changed the way I read and the way that I answer because I learned the skills that I need and what I need to do for every question (Interviewee 4).
>
> The teacher helped us improve reading a lot. I think now I'm better than before because now we learned many skills like scanning, highlight key words . . . stuff like that. This way of teaching changed how I read. We just following the steps and then we answer the question (Interviewee 7).

On entry to the college, students are placed into different levels according to the score on a government-administered English test taken at the end of high school. At the end of the foundation course, students take an *IELTS* Academic exam. Having different entry and exit exams means that we do not have an accurate measure of overall progress. The only general evaluation we have so far is from comparing separate (albeit similar) cohorts which is of limited value. However, it was interesting to note that *IELTS* Reading test scores at the end of the first year of implementing our reading model were not typical in that reading was not the weakest skill, as in previous years, but rather was the second strongest (after speaking).

We tried to develop a better understanding of the reading construct and the way our students read by investigating student responses to our tests. We conducted a small-scale study where we asked eight students to talk us through the process of answering questions in a test they had just done: a verbal reporting protocol. We wanted to find out how students did test tasks and to see if there was alignment between the cognitive processes that were used by students and those the test writer wanted to elicit. We found the process illuminating and instructive as we discovered how students approached the text and some of the skills and strategies they used, and we also uncovered flaws in the test design.

In terms of strategies used by students, we realised that higher level students seemed to be familiar with test strategies so we suggested that their teachers focus less on these and more on reading and developing vocabulary knowledge, with a greater emphasis on texts and less on question types. However, the lower level students we interviewed did not demonstrate good reading or test strategies and also reported problems with time, so teachers were asked to focus on time awareness and increasing reading speed as well as helping students with useful strategies. This experience made us realise that it is important to talk to our students about how they do reading tasks and encouraging them to compare and experiment with different strategies. We therefore asked teachers to devote some class time to group work and to having students discuss a text and answer questions together.

With some test questions we were able to find out if our tasks were actually testing what we aimed to test. For example a number of questions which the test writer had constructed to test students' ability to read and identify the main topic of a paragraph were not realised as students reported a strategy of matching words in the questions to words in the text, which resulted in the correct answer. This surface-structure reading was obviously much easier than the in-depth reading and inferencing predicted by the test writer but yielded the same correct answers. Students also reported misunderstanding badly-worded questions which resulted in them getting the answer wrong, and pointed out that some multiple-choice options could be eliminated through common sense without understanding the reading text. We realised after speaking to the students that our test covered expeditious and careful local reading very well but tended to neglect global reading, both careful (such as comprehending main ideas and overall text) and expeditious (such as skimming for gist). All of this information informed our test revision.

Conclusion

Our experience in developing a reading construct and using it in teaching and testing has added to our knowledge of L2

reading and has motivated us to explore reading further. When we interviewed students at the end of the first year, it was gratifying to listen to them talk confidently and knowledgeably about different reading skills and strategies and many students said they were more interested in reading and more willing to read in English as a result of their foundation year course. While we hope that our reading model will continue to motivate students to read and improve their reading skills, there are a number of limitations to our study that need to be recognised and addressed. We want to implement a more accurate measure of progress by using the same pre and post-test and one which can be aligned reliably to the CEFR. We are in the process of trialling different options with a view to using the test our institution chooses as a placement and exit test. We also recognise the need to provide more training for teachers and to involve teachers in all stages of the teaching and testing process. After talking to students from different classes, we realised that teachers had adopted the reading model to varying degrees and that there was some variety in coverage of reading skills and strategies between classes. We plan to meet teachers at the end of each semester to review practice and elicit feedback on all aspects of the reading model with a view to addressing any concerns, and we plan to involve more teachers in material and test design. We will also continue to use verbal reporting protocol to help us further understand students' reading and improve our tests.

Although this reading model was designed for our specific teaching context, we believe the process has general relevance. We would encourage teachers to consider strengthening alignment between skills teaching and testing by developing a construct and learning outcomes and initiating a cycle of evaluation and improvement. We are currently working with two colleagues who are adopting the same model for listening and it will be interesting to see how this unfolds.

References and further reading

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

Alderson, J C, Figueras, N, Kuijper, H, Nold, G, Takala, S and Tardieu, C (2004) *The development of specifications for item development and classification within the Common European Framework of Reference for Languages: learning, teaching, assessment. Reading and listening,* Final report of the Dutch CEF construct project, Lancaster: Lancaster University.

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

Fulcher, G (2010) *Practical Language Testing,* London: Hodder Education.

Gosling, D and Moon, J (2001) *How to Use Learning Outcomes and Assessment Criteria,* London: SEEC Office.

Grabe, W and Stoller, F (2002) *Teaching and Researching Reading,* London: Pearson Education.

IELTS: www.ielts.org/researchers/analysis_of_test_data.aspx

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

Kincaid, J P, Fishburne, R P, Rogers, R L and Chissom, B S (1975) *Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease formula) for Navy Enlisted Personnel,* Research Branch Report 8-75, Chief of Naval Technical Training: Naval Air Station Memphis.

Knowledge and Human Development Authority (2007) *TIMSS Educator's Report – Dubai 2007*, Dubai: KHDA.

Knowledge and Human Development Authority (2009) *Dubai PISA 2009 Report*, Dubai: KHDA.

Munby, J (1978) *Communicative Syllabus Design,* Cambridge: Cambridge University Press.

Nuttall, C (1996) *Teaching Reading Skills in a Foreign Language,* London: Heinemann.

Weir, C (2005) Limitations of the Common European Framework for developing comparable examinations and tests, *Language Testing* 22, 281–300.

Appendix A: Level 3 reading construct

Text details	
Type of texts	Authentic and adapted-authentic real-world notices; newspapers and magazines; simplified encyclopedias; brochures and leaflets; websites
Discourse types	Descriptive, narrative, instructive, expository, argumentative
Text purpose	Referential (to inform); poetic (to entertain); emotive (to convey feelings/emotions)
Nature of information	Concrete, opinion/feeling
Lexis	General vocabulary sufficient for most topics in everyday life
Grammatical structure	Simple sentences and a range of complex sentences
Cognitive demands	Lexis is familiar and structures mainly simple and easy to parse. Propositional meaning is quite low and inter-sentence relationships are quite simple
Text difficulty	The text types which can be handled by the learner at this level include street signs and public notices, product packaging, forms, posters, brochures, city guides and instructions on how to do things, as well as informal letters and newspaper and magazine texts such as articles and features A reading difficulty of up to 11, as measured by the Flesch-Kincaid readability index
Text length	600–800 words

Level 3/CEFR Level B1		
Type of reading	**Cognitive processes**	**Learning outcomes**
Overall reading comprehension	Processing will be at word, phrase, sentence, paragraph and whole text level	Can read and demonstrate comprehension of a selection of authentic and adapted-authentic texts including street signs and public notices, product packaging, forms, posters, brochures, city guides and instructions on how to do things, as well as informal letters, simplified encyclopedias, and newspaper and magazine texts such as articles and features. Individual texts will be up to 800 words in length with a reading difficulty between 9 and 11, as measured by the Flesch-Kincaid readability index.
Careful reading Local	Decoding level: *Word recognition* (spelling and pronunciation) *Lexical access* (form and meaning) *Syntactic parsing* (word order, word form, structural elements) *Establish core meaning* at clause and sentence level *Inferencing* at sentence level	Can read carefully to establish meaning at sentence level by identification of words and understanding structure of simple and a range of complex sentences. Can demonstrate comprehension by making inferences. Can read carefully to understand details.
Careful reading Global	*Building a mental model* Integration of information across longer stretches of text Follow text development within paragraph Comprehend main idea of paragraphs	Can demonstrate comprehension by deducing the meaning of unfamiliar words from context/form, and understanding discourse markers and grammatical cohesion. Can read carefully to establish main idea of paragraphs by understanding discourse markers, cohesion and organisation, and distinguishing main idea from supporting details. Can read carefully to establish purpose or main idea of the text by identifying or recognising main points that are central to the meaning. Can demonstrate comprehension of the writer's attitude or opinion.
Expeditious reading Local	*Scan* for specifics at word and clause level Recognition and matching word or group of words	Can scan text to look for specific words, phrases, names or specific items.
Expeditious reading Global	*Skim for gist* *Search reading*: Match words/words in similar semantic field Match sentences *Building a mental model* Integrating new information (Ongoing meaning representation of text) *Depending on experience of reader careful reading may take over when information has been quickly and selectively located*	Can skim text for general idea. Can search text for words to locate specific information. Can scan texts in order to locate desired information, and gather information from different parts of a text in order to fulfil a specific task.

Appendix B: Level 3 learning outcomes

Level 3/B1 Reading learning outcomes		**Weeks 1–3**	**Weeks 4–6**	**Weeks 7–9**	**Weeks 10–12**	**Weeks 13–15**	**Weeks 16–17**
01	Can read carefully to establish meaning at sentence level by identification of words and understanding structure of simple and a range of complex sentences.						
02	Can scan text to look for specific words, phrases, names or specific items.						
03	Can search text for words to locate specific information and read carefully to understand details.						
04	Can scan texts in order to locate desired information, and gather information from different parts of a text in order to fulfil a specific task.	■					
05	Can demonstrate comprehension by deducing the meaning of unfamiliar words from context/form.	■					
06	Can demonstrate comprehension by understanding discourse markers and grammatical cohesion.	■					
07	Can demonstrate comprehension by making inferences.	■	■				
08	Can read carefully to establish main idea of paragraphs by understanding discourse markers, cohesion and organisation, and distinguishing main idea from supporting details.	■	■				
09	Can skim text for general idea.	■	■	■	■		
10	Can read carefully to establish purpose or main idea of the text by identifying or recognising main points that are central to the meaning.	■	■	■	■		
11	Can demonstrate comprehension of the writer's attitude or opinion.	■	■	■	■		

ALTE report

ALTE 41st Meeting and Conference

Over 140 delegates attended ALTE's 41st bi-annual Meeting and Conference held at the Faculty of Letters of the University of Lisbon on 18–20 April 2012. The conference was hosted by CAPLE (Centro de Avaliação de Português Língua Estrangeira), the ALTE member for Portuguese and one of the founder members of the association.

The first two days of Workshops and Special Interest Group Meetings were attended by ALTE members and institutional affiliates, and the final day was an Open Conference Day for all those with an interest in language testing. The theme of the conference day was 'The Impact of Language Testing on Learning and Teaching' and considered issues related to understanding the relationship between learning, teaching and assessment. Researching this relationship is essential to ensuring that examinations help learners to achieve their life goals and have a positive impact on their learning and professional development.

After opening addresses from Professor António Feijó, Dean of the Faculty of Letters, and Dr Emyr Davies, Chair of the ALTE Executive Committee, Dr Waldemar Martyniuk, Executive Director of the European Centre for Modern Languages (ECML) in Graz, gave a brief overview of ECML's new programme of activities for the period 2012–15. This was followed by a presentation from Dr Dianne Wall, Trinity College London and Lancaster University, who talked about 'Examining Washback: What Do We Know, and What Is There Left To Explore?' Dr Wall provided a selective overview of research into washback and reviewed a number of studies that have helped practitioners not only to determine the effects of specific tests in specific contexts but also to expand our understanding of the original concept.

Professor Domingos Fernandes from the Faculty of Education, University of Lisbon, then discussed some of the theoretical and practical considerations associated with assessment. In his presentation entitled 'Assessment *for* Learning and Assessment *of* Learning', he argued that assessment can and must play a fundamental role in the teaching and learning processes and that assessing students' learning must be seen as integral to the process of teaching. Dr Nick Saville, Director Research and Validation, University of Cambridge ESOL Examinations then discussed the concept of 'Impact by Design' – an approach to impact research which examination providers such as ALTE members can use to find out and understand how their examinations impact on their stakeholders.

In his presentation 'Vocabulary in Language Proficiency Tests', Professor Norbert Schmitt of the University of Nottingham discussed ways of conceptualising lexical knowledge and how various item formats tap into different aspects of this knowledge. He also considered which item formats might be most appropriate for tests targeted at various proficiency levels.

José Pascoal from CAPLE, University of Lisbon, then looked at the 'Impact of The Common European Framework of Reference on Language Teaching and Assessment in Portugal' in his presentation and discussed issues related to the implementation of the CEFR in language policies in various educational contexts in Portugal. The conference closed with a presentation by the Director of CAPLE, Professor Maria José Grosso, who talked about the 'Impact of CIPLE (the Portuguese A2 examination) on Migration in Portugal'. Since 2006, immigrants seeking Portuguese citizenship have been required to show evidence of A2 level in Portuguese and Professor Grosso's presentation looked at how this has attracted a new group of candidates to the CIPLE examination.

Prior to the conference ALTE ran a two-day course on Basic Statistics for Language Testers. The course was presented by Michael Corrigan of the ALTE Validation Unit and Paul Crump of Assessment and Operations Group and looked at ways of using statistical information to facilitate test development and construction and to show how these techniques might fit into a test construction routine. ALTE also ran a one-day Foundation Course in Language Testing: Getting Started after the conference, which was run by Annie Broadhead, Consultant to Cambridge ESOL.

ALTE Language Testing Courses

ALTE is running two one-week-long Language Testing Courses at Hughes Hall in Cambridge in September. **The ALTE Introductory Course in Language Testing** will take place from 3–7 September and will be taught by Dr Lynda Taylor and Professor Cyril Weir of the University of Bedfordshire. **The ALTE Course in Understanding The C-levels to Assess Language for The Professions** will take place from 10–14 September and will be taught by Dr Anthony Green, University of Bedfordshire and Dr Fiona Barker, University of Cambridge ESOL Examinations.

ALTE 42nd Meeting and Conference

ALTE will hold its 42nd bi-annual Meeting and Conference at the headquarters of the Goethe-Institut in Munich on 21–23 November 2012.

As with previous events, the first two days of meetings will be for representatives of ALTE members and institutional affiliates only, and the final day, Friday 23 November, will be an open conference day for all those with an interest in language testing. The theme of the conference day will be 'Developing and Adapting Test Materials for Younger Learners' and speakers will include Shelagh Rixon of the University of Warwick, Henny Rönneper from the Ministry of Education and School, Nordrhein-Westfalen, and Neil Jones of University of Cambridge ESOL Examinations. A number of representatives of the SurveyLang project will also give their perspectives on aspects of the project.

Prior to the Meeting and Conference ALTE will run a two-day Introductory Course on Assessing Young Learners on 19–20 November, and a one-day Foundation Course on Language Testing: Getting Started on 20 November.

For further information about all ALTE activities, please visit the ALTE website – www.alte.org.

Reader Survey

Research Notes has reached its twelfth year in publication and we are interested in your views about its content and approach.

Please take the time to complete this short survey as your responses will help inform the future development of this publication and will provide Cambridge ESOL with a clearer picture of the needs and interests of its readers.

This survey can be completed online at: www.surveymonkey.com/s/Research_Notes_Readership_Survey

Send your completed survey to:
John Savage
Research and Validation Group
University of Cambridge ESOL Examinations
1 Hills Road
Cambridge
CB1 2EU
United Kingdom
email: validation@cambridgeesol.org
Fax: +44 (0) 1223 553083

1. What is your main occupation?

☐ Lecturer ☐ Researcher ☐ Teacher
☐ Administrator ☐ Language tester ☐ Student
Other (please specify)

.....

2. How long have you been involved in language education?

☐ Less than 1 year ☐ 1–5 years ☐ 6–10 years
☐ 11–15 years ☐ More than 15 years
☐ Not applicable

3. How long have you been involved in language testing?

☐ Less than 1 year ☐ 1–5 years ☐ 6–10 years
☐ 11–15 years ☐ More than 15 years
☐ Not applicable

4. How did you find out about Research Notes?

☐ Direct mailing ☐ Cambridge ESOL website
☐ Conference ☐ Promotional presentation
☐ Referenced in another publication
☐ From an instructor/colleague
Other (please specify)

.....

5. Research Notes is published 4 times per year. How many issues do you usually read (in whole or in part) per year?

☐ 4 issues ☐ 3 issues ☐ 2 issues ☐ 1 issue
☐ None

6. How much of each issue do you usually read?

☐ All of it ☐ Most of it ☐ Some of it
☐ None of it

7. Have you accessed further information from Cambridge ESOL after reading Research Notes?

☐ Yes (go to question 8) ☐ No (go to question 10)

8. What information did you access?

.....

.....

.....

.....

9. Was it easy to find the information you were looking for?

☐ Yes ☐ No
Comments:

.....

.....

.....

.....

Research Notes publishes a variety of articles on a range of topics. The following 3 questions focus on which topics you are most interested in reading.

10. Which exams are you most interested in reading about?

	Very interested in	Somewhat interested in	Not very interested in
Young learners and for Schools exams	☐	☐	☐
Academic and Professional English exams	☐	☐	☐
General English exams	☐	☐	☐
Business English exams	☐	☐	☐
Cambridge English Teaching Qualifications	☐	☐	☐

Other (please specify)

.....

.....

11. Which skills/systems are you most interested in reading about?

	Very interested in	Somewhat interested in	Not very interested in
Listening	☐	☐	☐
Speaking	☐	☐	☐
Reading	☐	☐	☐
Writing	☐	☐	☐
Grammar	☐	☐	☐
Vocabulary	☐	☐	☐

Other (please specify)

.....

.....

12. Which topics are you most interested in reading about?

	Very interested in	Somewhat interested in	Not very interested in
Teaching/learning	☐	☐	☐
Content and Language Integrated Learning (CLIL)	☐	☐	☐
Research methods	☐	☐	☐
Validity/reliability	☐	☐	☐
Washback/impact	☐	☐	☐
Test design/development	☐	☐	☐
Technological developments in testing	☐	☐	☐
Standard setting/ benchmarking	☐	☐	☐
Common European Framework of Reference (CEFR)	☐	☐	☐
English Profile	☐	☐	☐
Using corpora	☐	☐	☐
Operational processes of an exam board (examiner training, etc.)	☐	☐	☐
Learning oriented assessment/formative assessment	☐	☐	☐
Conference/publication updates	☐	☐	☐

Other (please specify)

.....

.....

.....

.....

13. How would you prefer to receive Research Notes?

☐ a link via email ☐ a hard copy via post

14. Do you have any other comments on Research Notes?

.....

.....

.....

.....

1770222649

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

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

RESEARCH NOTES

Contents:

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

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

ALTE Association of Language Testers in Europe

PART OF THE CAMBRIDGE ASSESSMENT GROUP